Lending Club Case Study

By:

Vikas Sunkad & Venkat Chalama Reddy

Index

1. **Broad Overview of Case Study Understanding**
    - Problem Statement
    - Analysis Approach
2. **Data Understanding**
3. **Data Pre-processing**
4. **Exploratory Data Analysis**
    - Univariate Analysis
    - Segmented Univariate Analysis
    - Bivariate Analysis
5. **Actionable Insights**

Broad Overview of Case Understanding

Problem Statement

- A Consumer finance company, specializes in lending various types of loans to urban customers (personal loans, business loans and financing of medical procedures), want's to build a risk analytics process of evaluating a borrower's ability to repay the loan and determine the likelihood of default.

- **The goal is to identify strong predictors of loan default (charged-off status).**

- **Output:** A model or insight into which factors are most indicative of risk, helping improve loan approval processes and minimize credit loss.

Business Understanding

- A Consumer finance company, when receives a loan application, has to make a decision for loan approval based on the applicant’s profile. As part of risk mitigation, company want's to reduce the business and financial or Credit loss. During evaluating a loan application there are two types of risks associated with the bank’s decision:
  1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
  2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Business Objective

- Company's decision making in loan approval process is backed with strong analytics data, which helps to take a right decision for Loan Sanction.
- Company's loan approval time can be reduced, which improves profitability and business growth.
- Company can mitigate the credit loss with help of risk analytics, by identifying the borrowers capability of loan repayment and rejecting the loan application.
- Company can mitigate the business loss with help of risk analytics, by not rejecting a loan to a customer whose loan repayment status and credit score is very good or exceptional.
- Company can take a decision on interest rate provided to customers based on the risk appetite i.e. sanction a loan with high interest rate to a customer with average credit score and low interest rate to a customer with exceptional credit score.

Approach

- **Data Understanding**
    - Examine the key variables in the dataset and build an understanding of what each means.

- **Data Preprocessing**
    - Handle missing data, remove irrelevant features, outlier detection and treatment, feature engineering

- **Exploratory Data Analysis (EDA)**
    - Univariate analysis, Segmented Univariate Analysis, Bivariate Analysis

- **Actionable Insights**

Data Understanding

Loan File

- The Loan.csv file has 39717 records
- There are a total of 111 columns
- The columns can be broadly be classified as follows:

Classification	Data Columns
Identifier	Id, member_id
Loan Details	loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, issue_d
Person Details	emp_title, emp_length, home_ownership, annual_inc, url, zip_code, addr_state, application_type
Control Details	verification_status, pymnt_plan, policy_code
Financial Details	desc, purpose, title, dti, delinq_2yrs, earliest_cr_line, inq_last_6mths, mths_since_last_delinq, mths_since_last_record, open_acc, pub_rec, revol_bal, revol_util, total_acc, initial_list_status, pub_rec_bankruptcies
Recovery Details	out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, next_pymnt_d, last_credit_pull_d
Does not contain Data	59 columns which do not contain any data can be dropped from the analysis
Output Details	loan_status

Loan File – Does not contain Data

Classification	Data Columns
Does not contain Data	collections_12_mths_ex_med, mths_since_last_major_derog, annual_inc_joint dti_joint, verification_status_joint, acc_now_delinq, tot_coll_amt, tot_cur_bal open_acc_6m, open_il_6m, open_il_12m, open_il_24m, mths_since_rcnt_il, total_bal_il, il_util, open_rv_12m, open_rv_24m, max_bal_bc, all_util, total_rev_hi_lim, inq_fi, total_cu_tl, inq_last_12m, acc_open_past_24mths, avg_cur_bal, bc_open_to_buy, bc_util, chargeoff_within_12_mths, delinq_amnt mo_sin_old_il_acct, mo_sin_old_rev_tl_op, mo_sin_rcnt_rev_tl_op, mo_sin_rcnt_tl mort_acc, mths_since_recent_bc, mths_since_recent_bc_dlq, mths_since_recent_inq, mths_since_recent_revol_delinq, num_accts_ever_120_pd, num_actv_bc_tl, num_actv_rev_tl, num_bc_sats, num_bc_tl, num_il_tl, num_op_rev_tl, num_rev_accts, num_rev_tl_bal_gt_0, num_sats, num_tl_120dpd_2m, num_tl_30dpd, num_tl_90g_dpd_24m, num_tl_op_past_12m, pct_tl_nvr_dlq, percent_bc_gt_75, tax_liens, tot_hi_cred_lim, total_bal_ex_mort, total_bc_limit, total_il_high_credit_limit

Understanding Data in more detail…

Classification	Data Columns	Impact on Output Variable
Identifier	Id, member_id, url	These will not have any impact on the output variable. So, no further analysis will be done on this data
Loan Details	funded_amnt	We will have to study this column a bit more in detail, to understand whether higher loan amount can translate to higher risk of default or any other interplay.
	int_rate	We will have to study the impact of this variable as well. If interest burden increases on people, are they likely to default. This needs to be studied.
	grade, sub_grade	These two variables talk about the grade i.e. quality score. A lower grade should indicate higher risk and hence needs to be studied more in detail.
	Remaining columns	No impact. Not to be studied in any further detail.
Person Details	emp_length, home_ownership, annual_inc	More the length of an employee lower is the risk of default. A homeowner is also less likely to default. Similarly a person with high income is less likely to default. Hence these need detailed analysis.
	Remaining columns	No impact. Not to be studied in any further detail.
Control Details	verification_status	A verified income is a more reliable data vs. unverified income. Hence the possibility of unverified income to be source of risk needs to be studied.
	Remaining columns	No impact. Not to be studied in any further detail.

Understanding Data in more detail…

Classification	Data Columns	Impact on Output Variable
Financial Details	Purpose	This indicates the purpose for which the loan is being applied for. We need to study whether the certain types of loans are inherently more risky.
	dti	This is an important ratio and indicates how much is the monthly payout Vs. the Salary. Lower the ratio indicates a better financial position and hence possibly lower risk.
	delinq_2yrs	Indicates the number of delinquencies in the last two years. Higher this figure, higher the risk of default.
	inq_last_6mths	Indicates the number of times a borrower has applied for loan in the last 6 months. Higher the no. of enquiries one could construe that the borrower is possibly in financial distress and hence can be thought of as risky.
	open_acc	This includes all active credit lines such as credit cards, installment loans, retail accounts, and other revolving or non-revolving credit accounts. Very few and very high open accounts can indicate risky behavior, while a healthy number of accounts can indicate moderate to low risk.
	pub_rec	This field indicates the number of severe or derogatory public records in a borrowers file. A figure of ‘0’ indicates a clean historical record, while a figure of 1 or more indicates significant financial event

Understanding Data in more detail…

Classification	Data Columns	Impact on Output Variable
Financial Details	revol_util	The percentage of available revolving credit that the borrower is currently utilizing. Borrowers with low revol_util have a good credit rating score and are less likely to default and vice versa.
	pub_rec_bankruptcies	A subset of pub_rec, but focuses specifically on bankruptcies, which are very severe form of financial events.
	mths_since_last_delinq	Indicates financial delinquency behavior. A figure of NA indicates that the borrower has never had a financially delinquent behavior in their credit history. A low figure indicates that the incident happened very recently and can be an indicator of risk, while a big figure tells us how much time back such episode happened. This is related to "delinq_2yrs" field.
	mths_since_last_record	This is related to "pub_rec" field.
Recovery Details	All columns	Does not have a direct link to the assessment of the risk profile of the customer, as this after the loan has been issued.

Data Preprocessing

Actions taken

- **Finding Missing Values:**
  - Drop the columns whose entire dataset is NULL.
- **Check for Duplication**
- **Check for Unique values**
  - Drop columns with single unique value and doesn't have a predictive power to identify the customer is defaulted or not.
- **Data Reduction**
  - Drop the columns in the dataset, which doesn't have a predictive power to identify the customer is defaulted or not .
- **Imputing Values**
  - Where the variable does not have any impact on charged off % and hence we are ignoring the missing values (emp_length).
  - Imputed with zero where we could correspond with another variable (revol_util NaN is replaced with 0 where revol_bal is 0 & pub_rec_brankruptices NaN is replaced with 0 where pub_rec is 0).
  - In other cases the rows with NaN values have been deleted.
- **Handling incorrect data type**
  - Numerical data with % have been converted float by removing % (e.g. int_rate, revol_util)
  - Some float data has been converted into int (e.g. pub_rec_bankruptcies, annual_inc)

Actions taken

- **Sanity Check**
  - Converting categorical variable like emp_length, term to a numerical variable.
  - Correcting some incorrect classification like “NONE” in home_ownership to “OTHER”.
  - Removing the rows with loan_status as “Current”.
- **Creating Utility Functions:**
  - For Box Plots, Calculating percentages which are used repeatedly in the analysis.
- **Outlier Analysis**

Exploratory Data Analysis

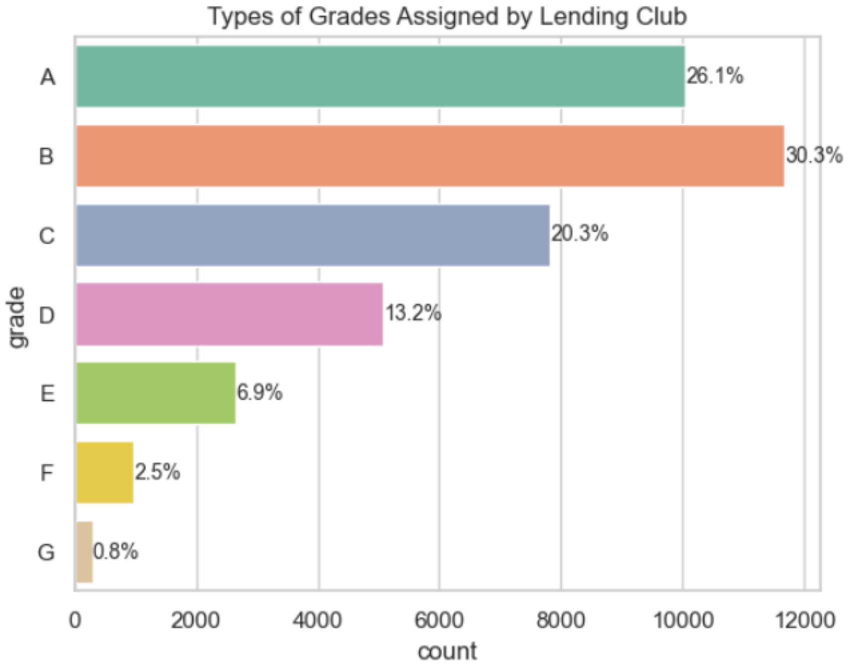
Univariate Variables

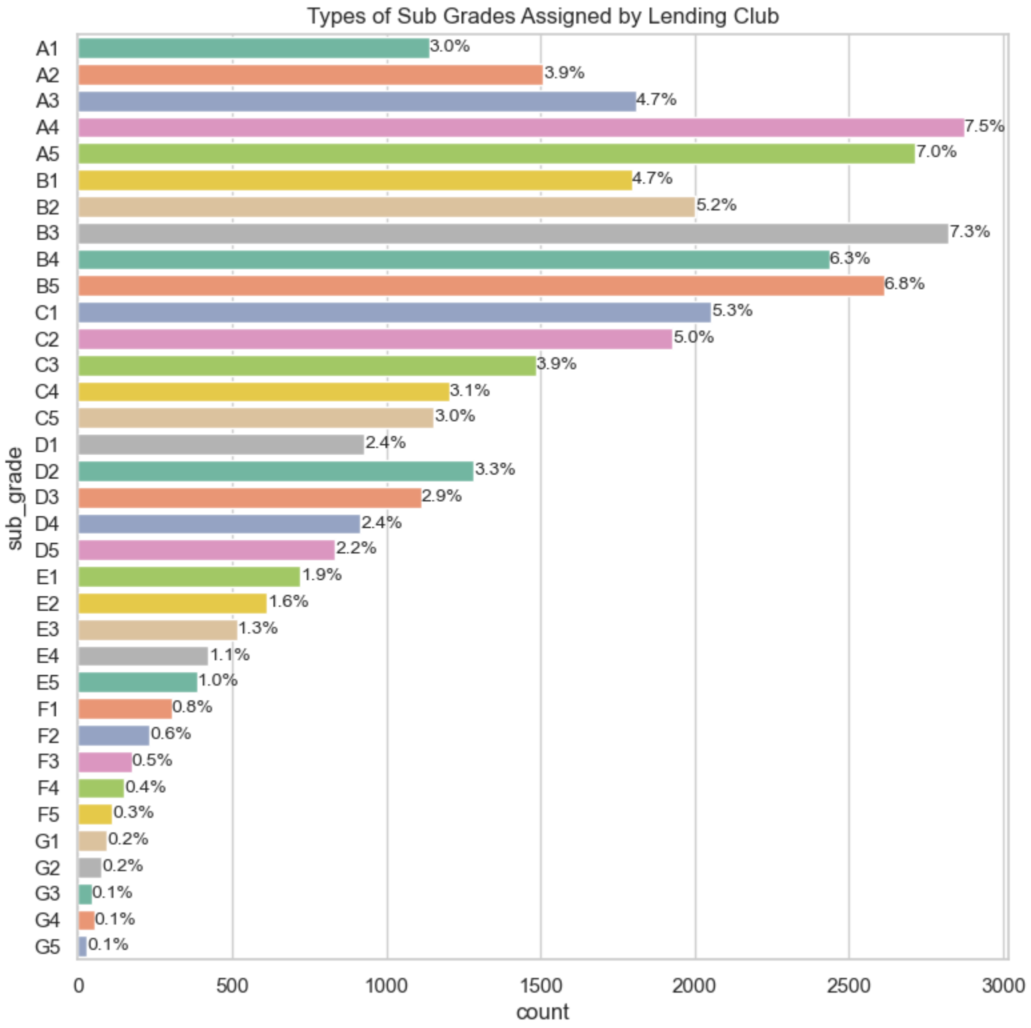
Variables used in univariate analysis are categorized as below:

1. **Categorical Variables Ordered Categorical:**
   - Grade (grade)
   - Sub-grade (sub_grade)
2. **Unordered Categorical:**
   - Home Ownership (home_ownership)
   - Loan purpose (purpose)
   - Loan Verification Status (verification_status)
3. **Quantitative Variables:**
   - Loam Amount (loan_amnt)
   - Interest Rate (int_rate)
   - Installment (installment)
   - Annual Income (annual_inc)
   - Earliest Credit Line (earliest_cr_line)
   - Debt to Income Ratio (dti)
   - No of Enquires for last 6 mnths (inq_last_6mths)
   - Open Credit lines (open_acc)
   - Revolving Balance (revol_bal) - outstanding amount of money you owe on credit cards or other forms of revolving credit.
   - Revolving Util (revol_util) - percentage of your available credit that you're currently using on revolving accounts (primarily credit cards).
   - Employee length (emp_length)
   - Loan term (term - 36, 60)

Ordered Categorical Variables

- **Grade & Sub Grade**





Grade:

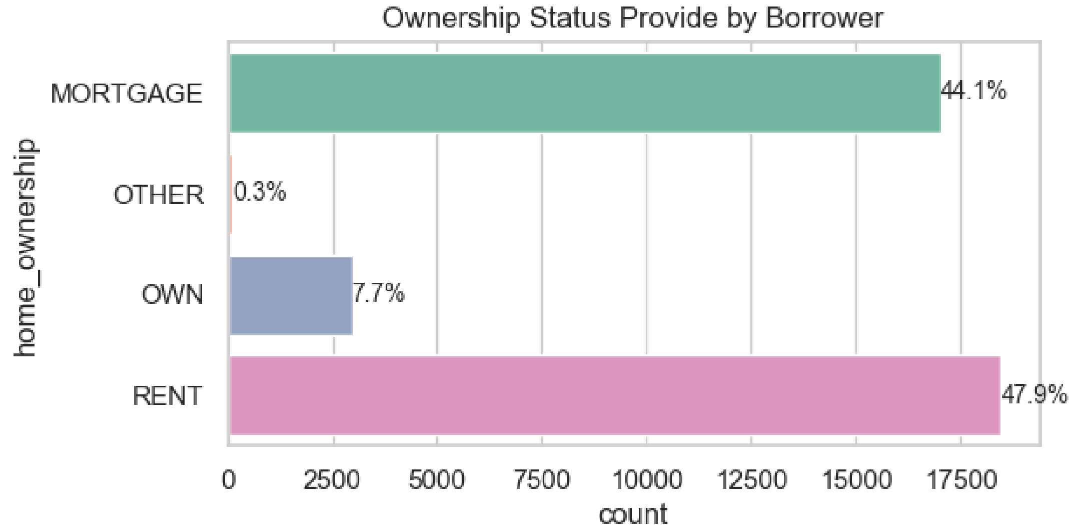
- 30.3% loans provided to borrowers with Grade B.
- 26.1% loans provided to borrowers with Grade A.
- 20.3% loans provided to borrowers with Grade C.

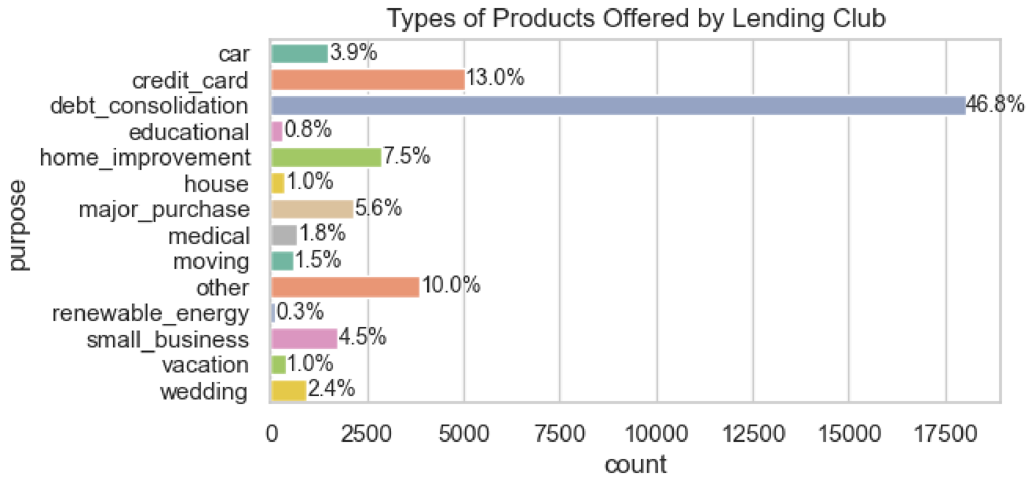
Sub Grade:

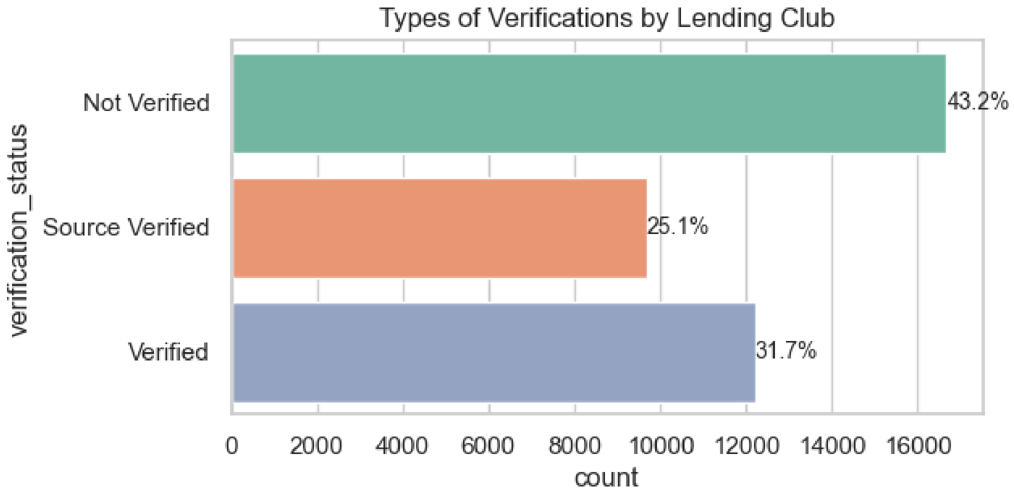
- 7.5% loans provided to borrowers with Sub Grade A4.
- 7.3% loans provided to borrowers with Sub Grade B3.
- 7.0% loans provided to borrowers with Sub Grade A5.

Unordered Categorical Variables

- **Home Ownership, Purpose & Verification Status**







Home Ownership

- 47.9% of borrowers are Renters
- 44.1% of borrowers are Mortgage
- 7.7% of borrowers are Home Owners

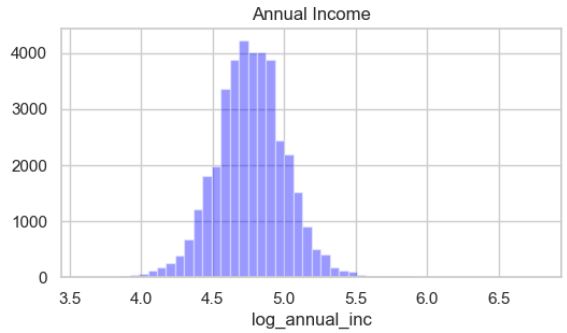
Purpose

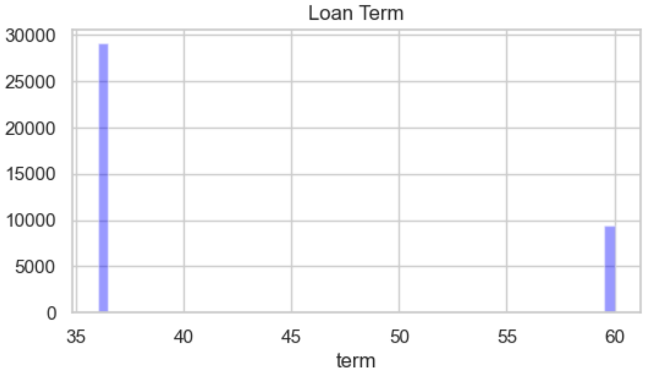
- 46.8% are Debt consolidation loans
- 13.0% are Credit Card loans
- 10.0% are Other purpose loans
- 7.5% are Home Improvement loans
- 5.6% are Major purpose loans

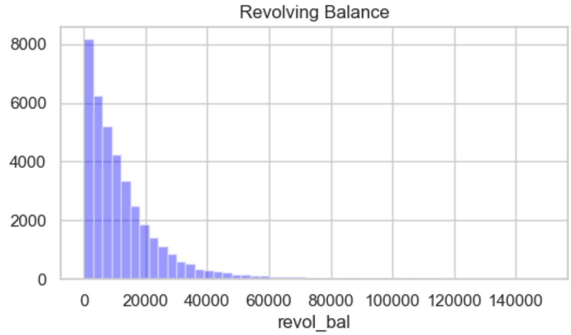
Verification Status

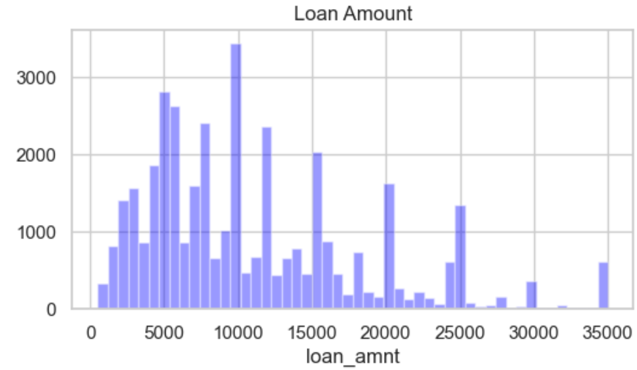
- 43.2% of loans are Not Verified
- 31.7% of loans are Verified
- 25.1% of loans are Source Verified

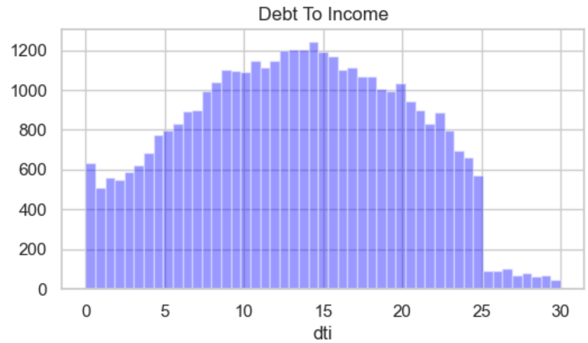
Quantitative Variables

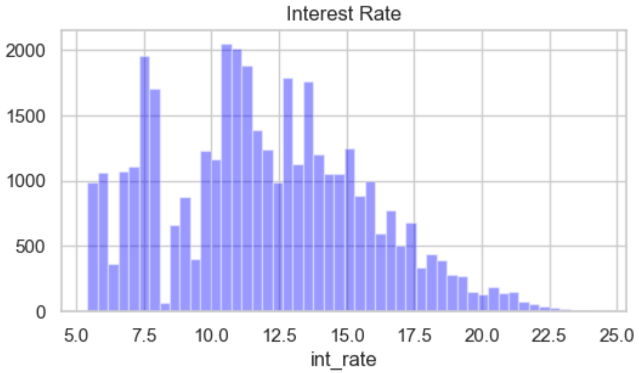


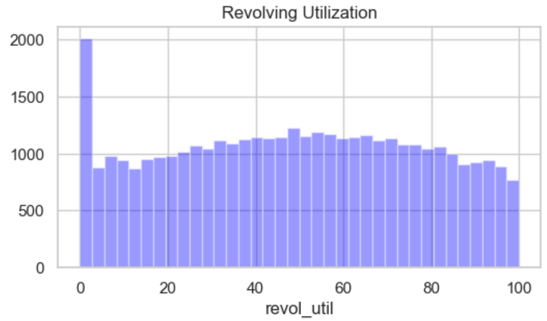


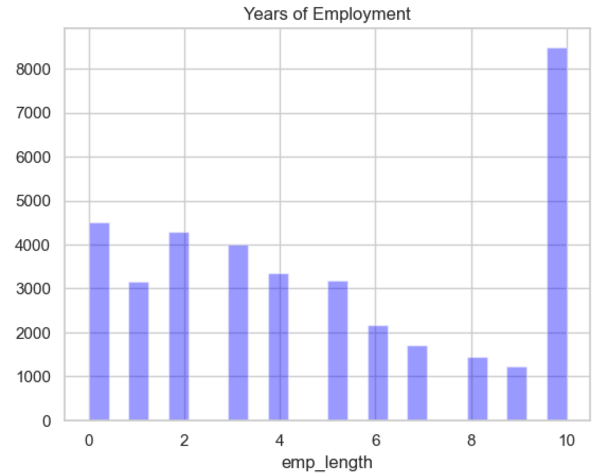


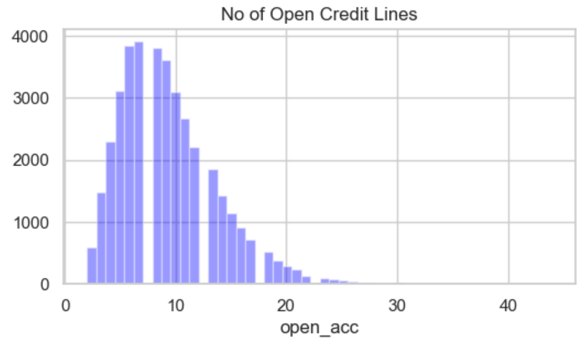


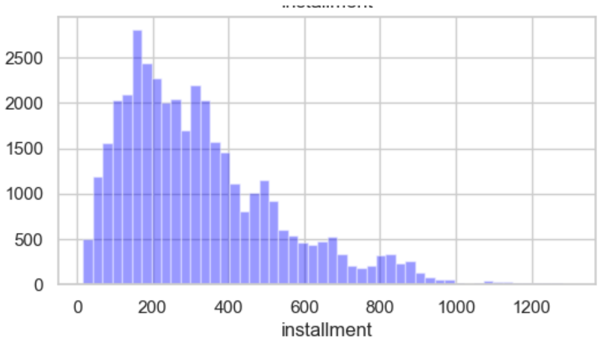


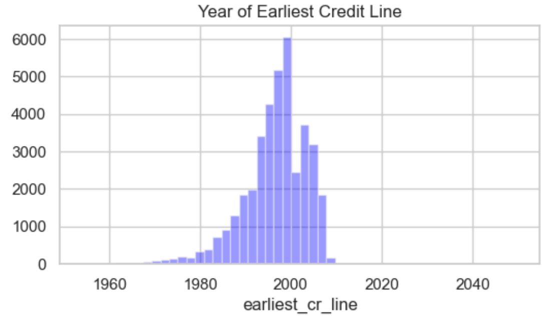








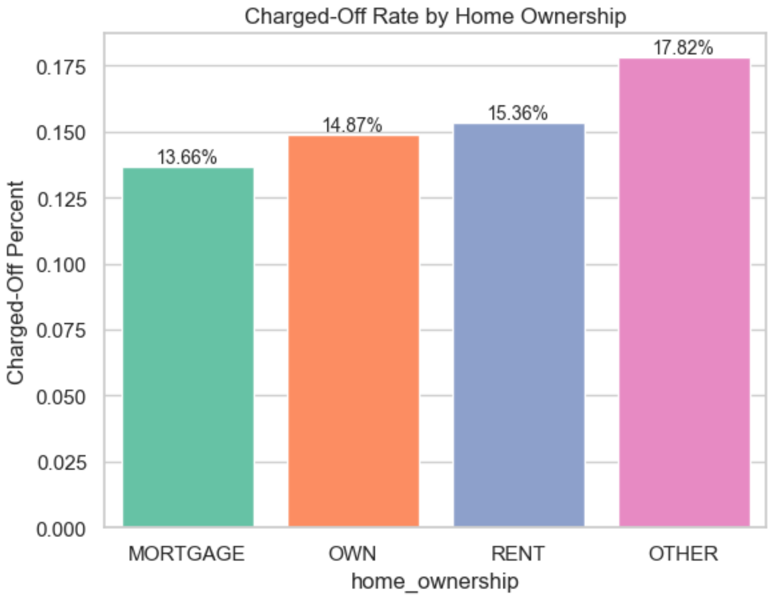


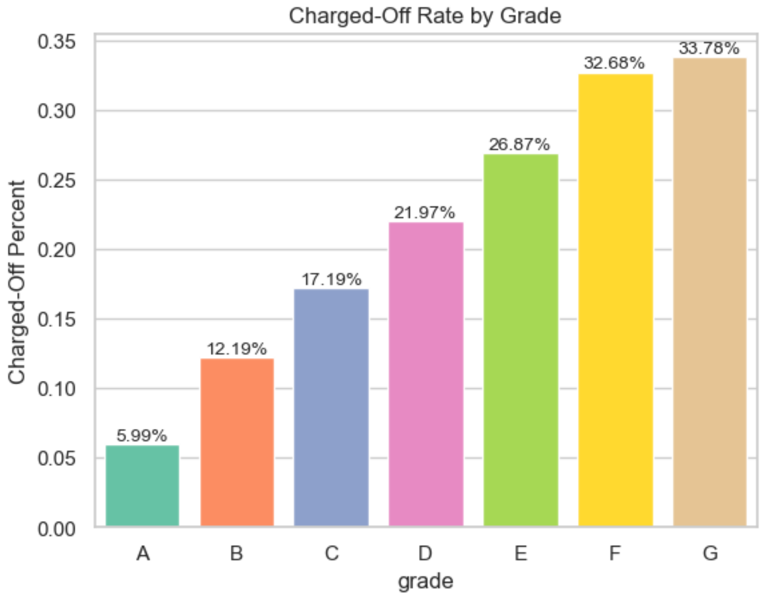


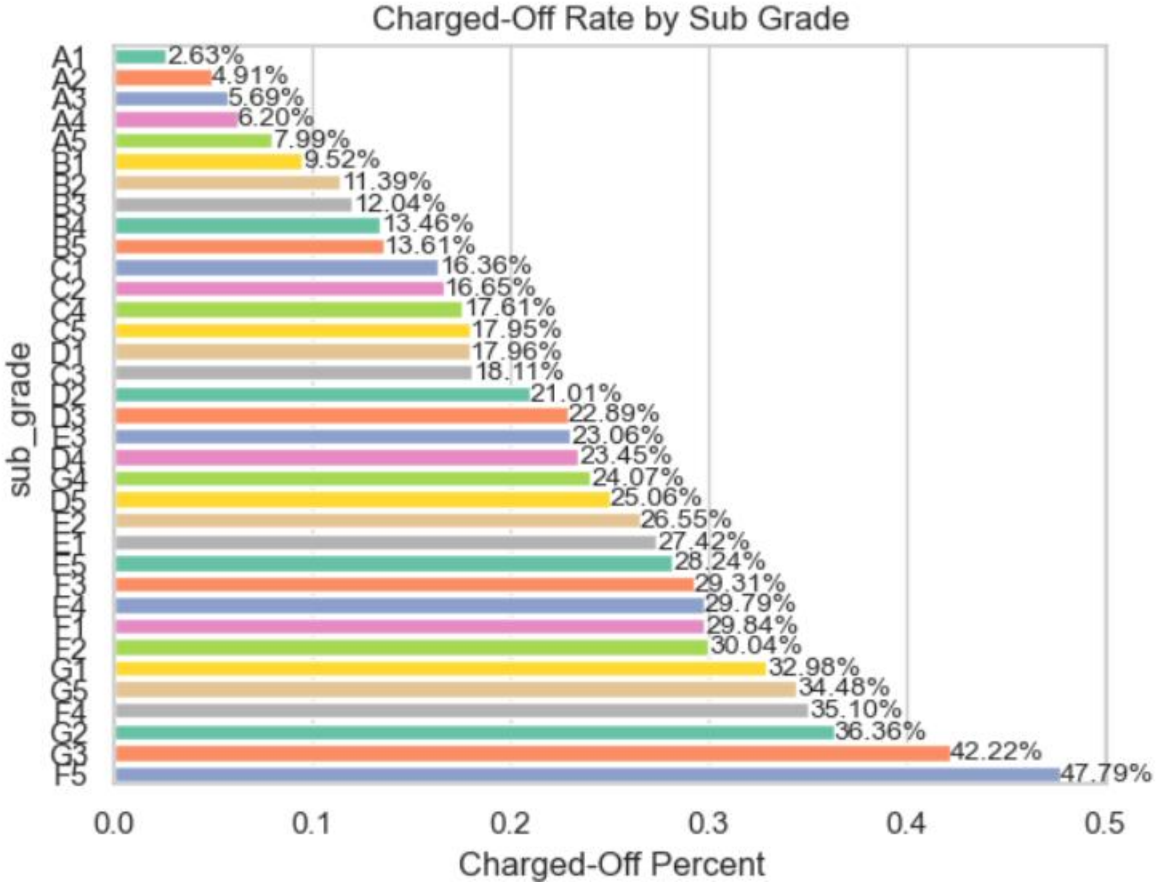
Segmented Univariate Analysis

1. **Segmented Univariate Analysis :**
    - Loan purpose (purpose)
    - Home Ownership (home_ownership)
    - Sub-grade (sub_grade)
    - Grade (grade)
    - Loan Verification Status (verification_status)

Segmented Univariate Variables

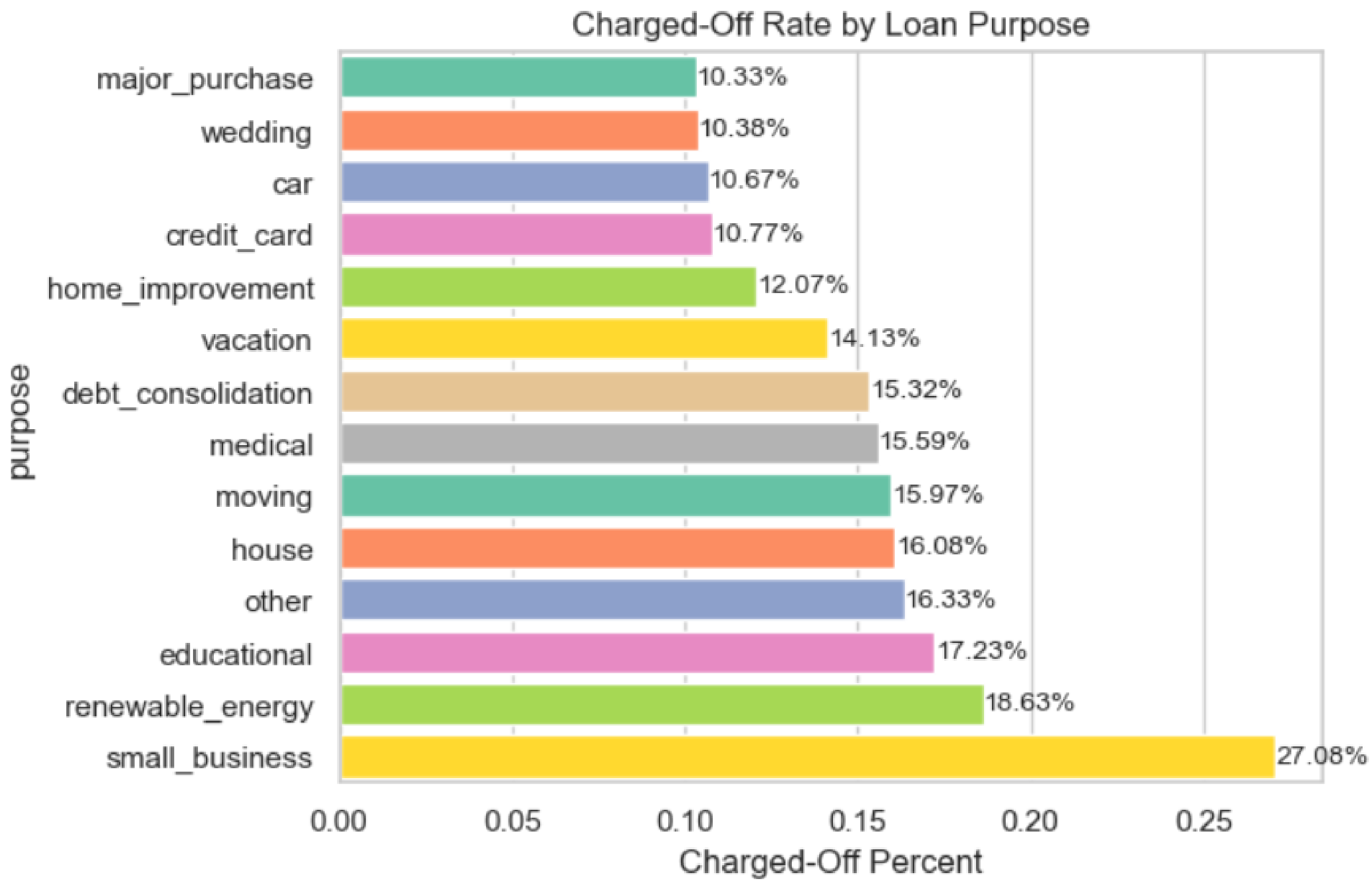


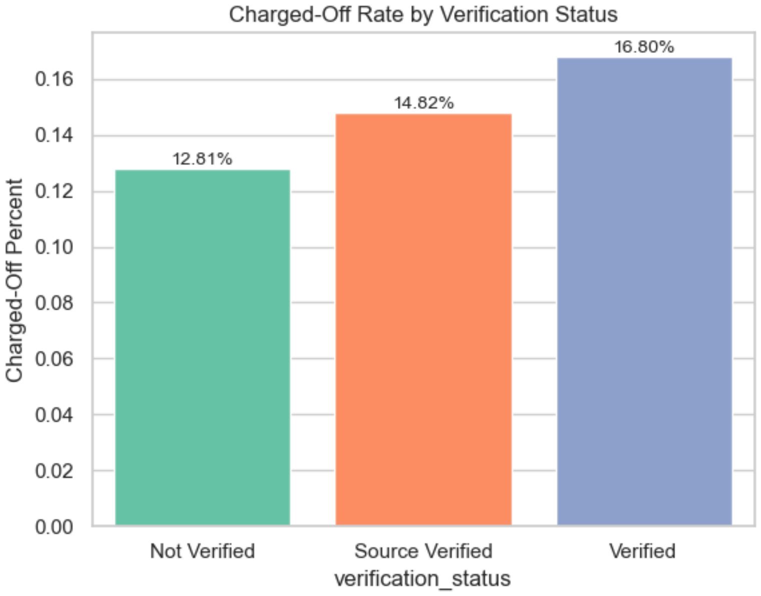




- **Home Ownership** - Charged-off loans are higher for Home Ownership Other i.e. **17.82%**
- **Grade** - Charged-off loans are highest for **Grade - G** i.e. **33.78%**
- **Sub Grade** - Charged-off loans are highest for **Sub Grade – F5** i.e. **47.79%**

Segmented Univariate Variables…


Charged-Off Rate by Loan Purpose

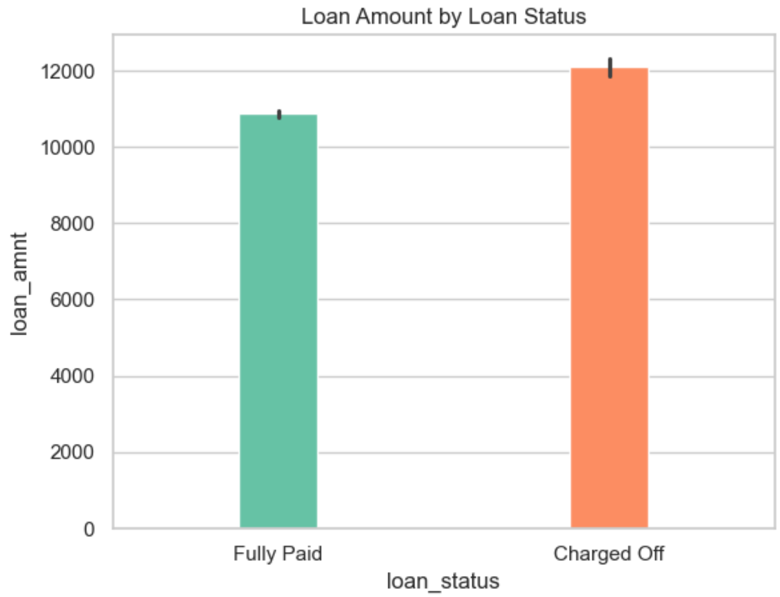

Charged-Off Rate by Verification Status

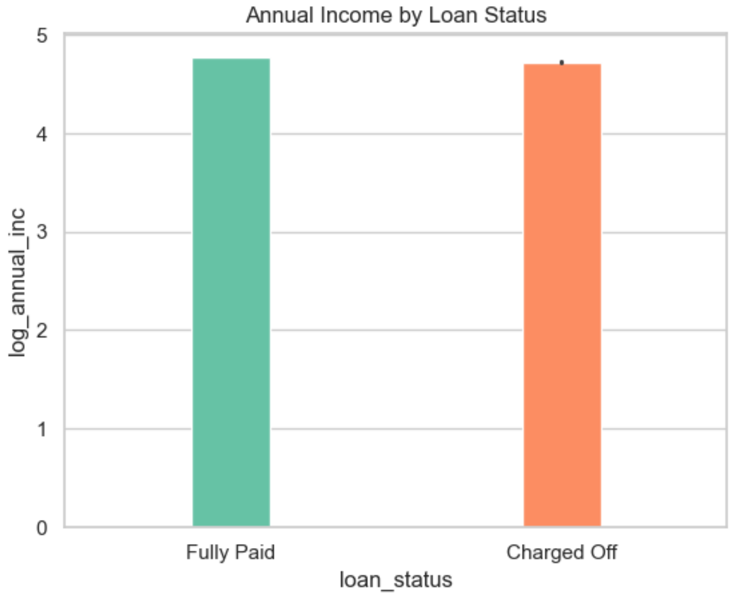
- **Loan Purpose** - Charged-off loans are higher for Small business i.e. **27.8%**
- **Verification Status** - Charged-off loans are higher for Verified Borrowers i.e. **16.80%**

Bivariate Analysis

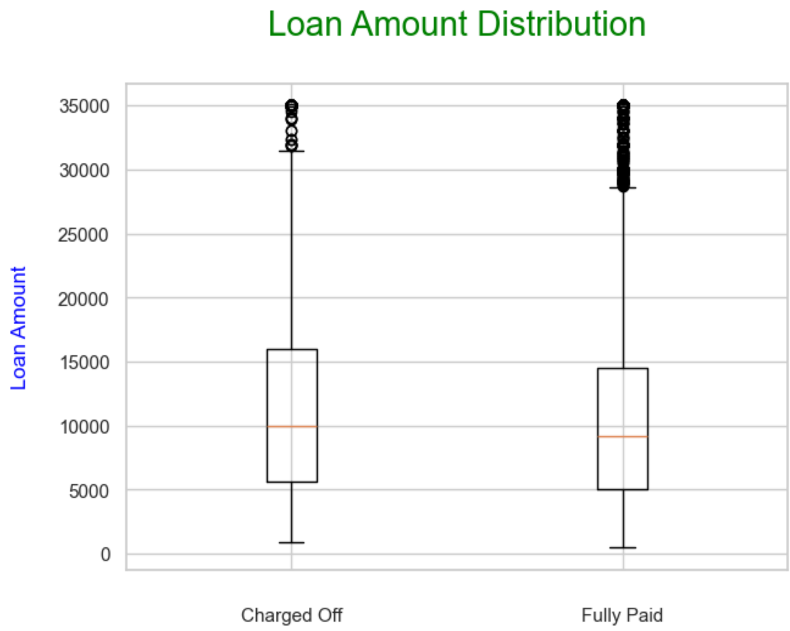
- **Continuous Variables**
  - Loam Amount (loan_amnt)
  - Interest Rate (int_rate)
  - Installment (installment)
  - Annual Income (annual_inc)
  - Earliest Credit Line (earliest_cr_line)
  - Debt to Income Ratio (dti)
  - No of Enquires for last 6 mnths (inq_last_6mths)
  - Open Credit lines (open_acc)
  - Revolving Balance (revol_bal) - outstanding amount of money you owe on credit cards or other forms of revolving credit.
  - Revolving Balance (revol_util) - percentage of your available credit that you're currently using on revolving accounts (primarily credit cards).
  - Loan term (term - 36, 60)
  - Public Records to Assess Creditworthiness(pub_rec)

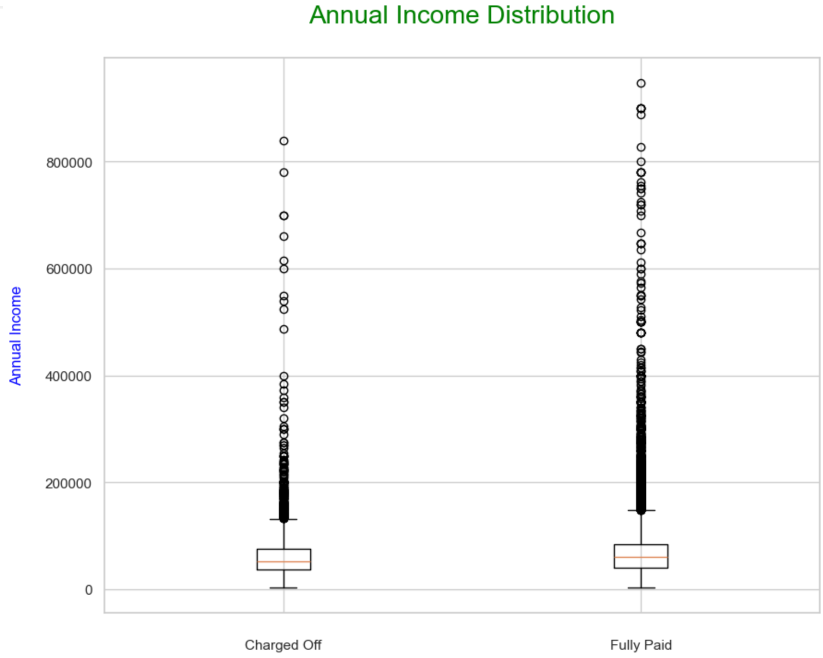
Continuous Variables



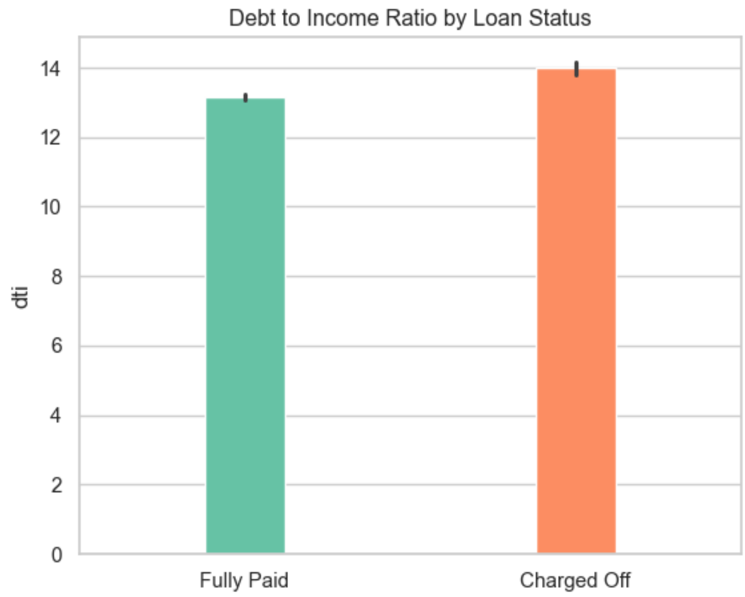


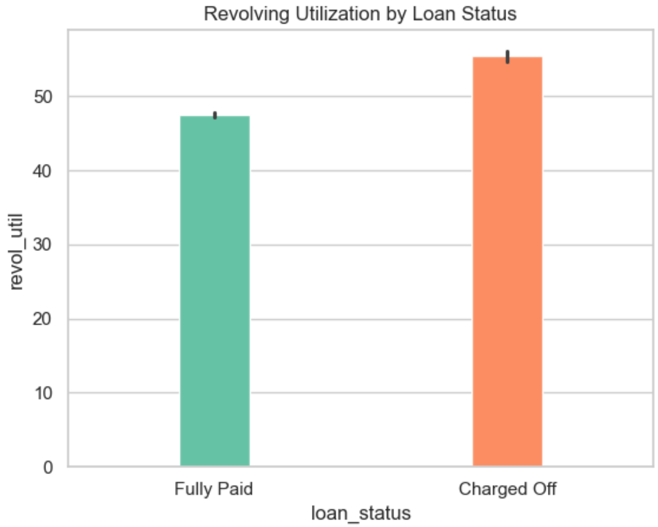
Loan Amount & Annual Income are not influencing the Charged off loans.



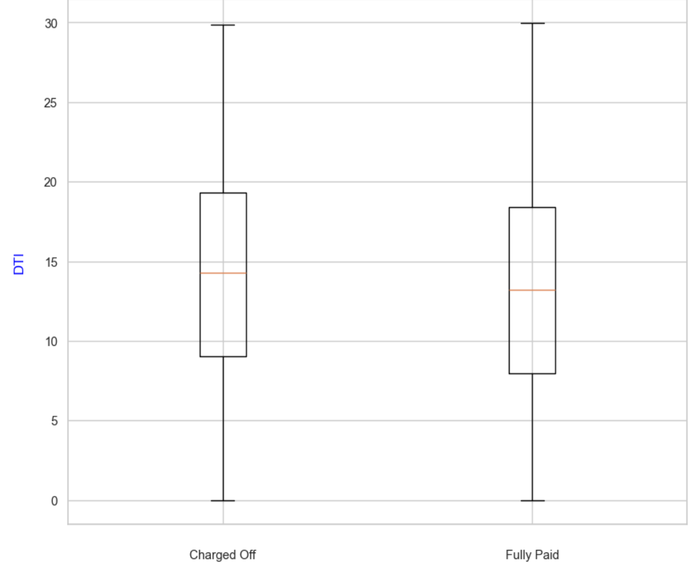


Continuous Variables

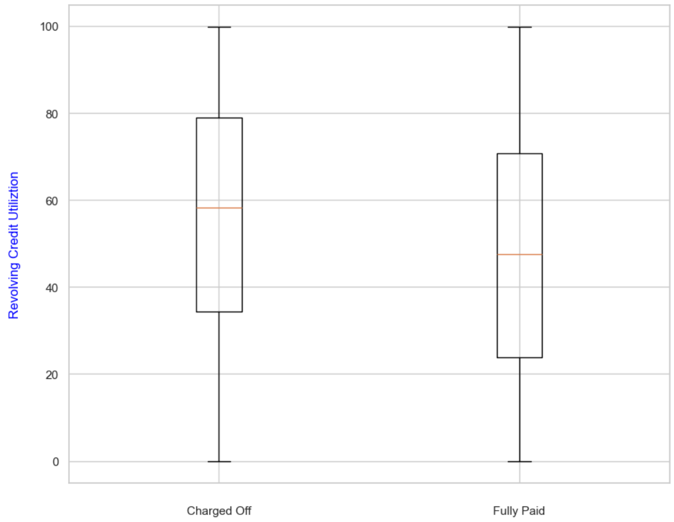

Debt to Income Ratio by Loan Status


Revolving Utilization by Loan Status

DTI Distribution



Revolving Credit Utilization Vs. Charge Off

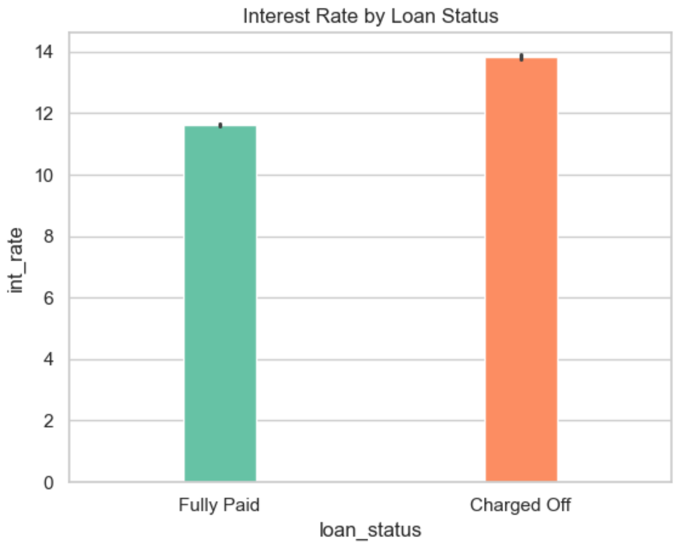


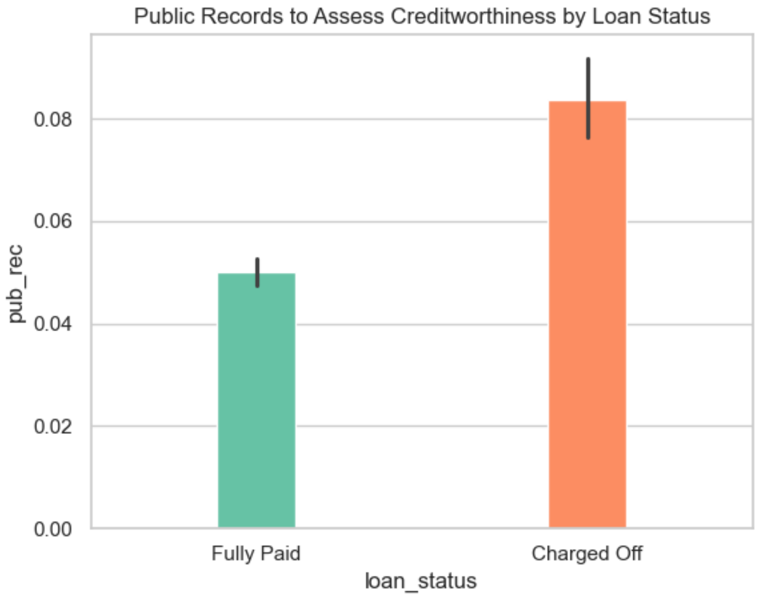
- Revolving utilization has higher Charged-off percentage.
- DTI doesn't have any major impact On Charged-off percentage.

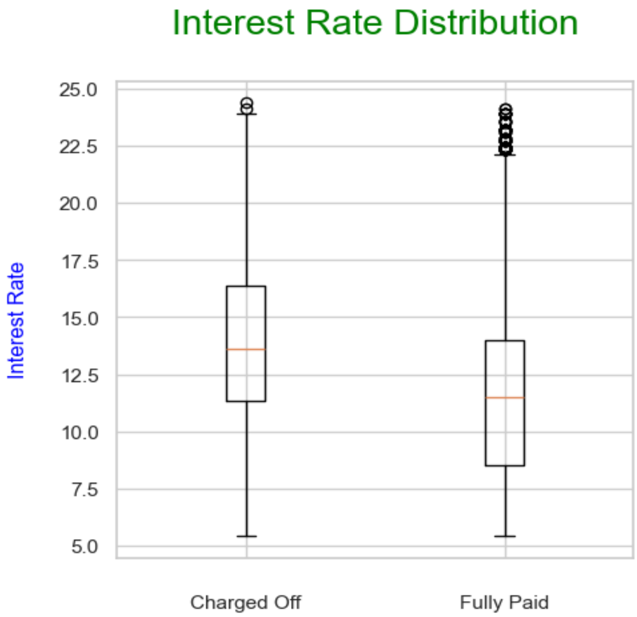
Continuous Variables

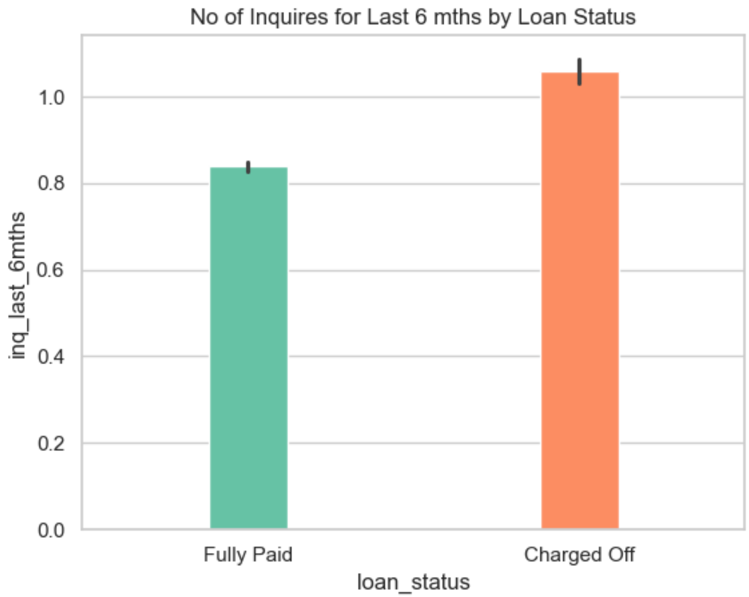
- Interest rates are higher for Charged-off loans.

- Higher number of public records and No of inquiries last 6 months results in higher Charged-off percentage.

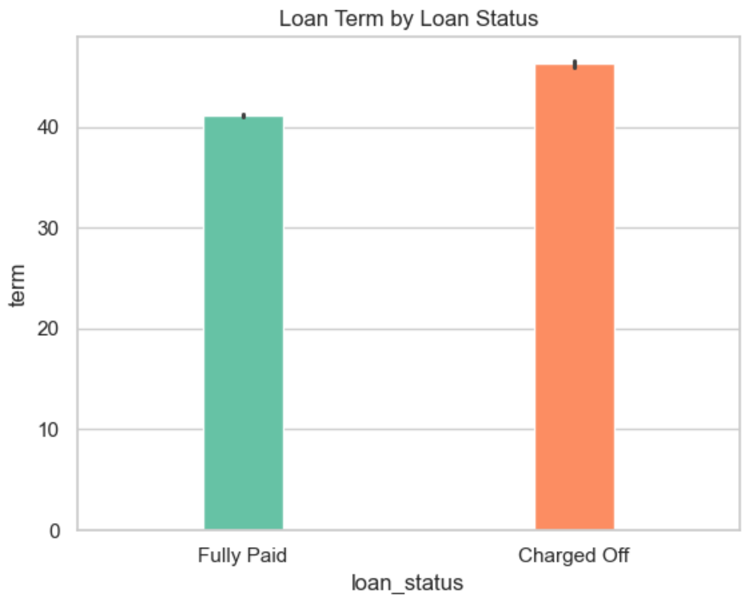


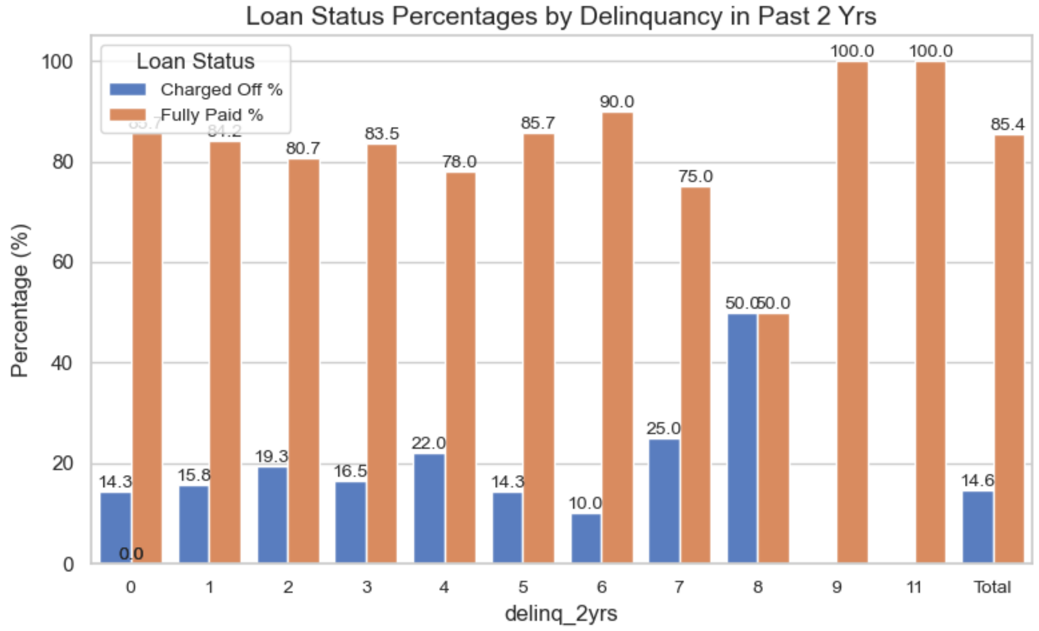


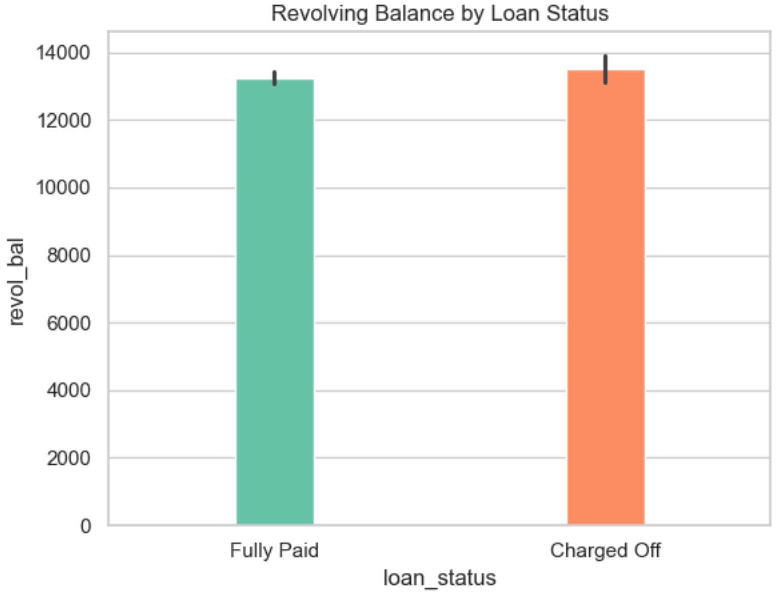




Continuous Variables

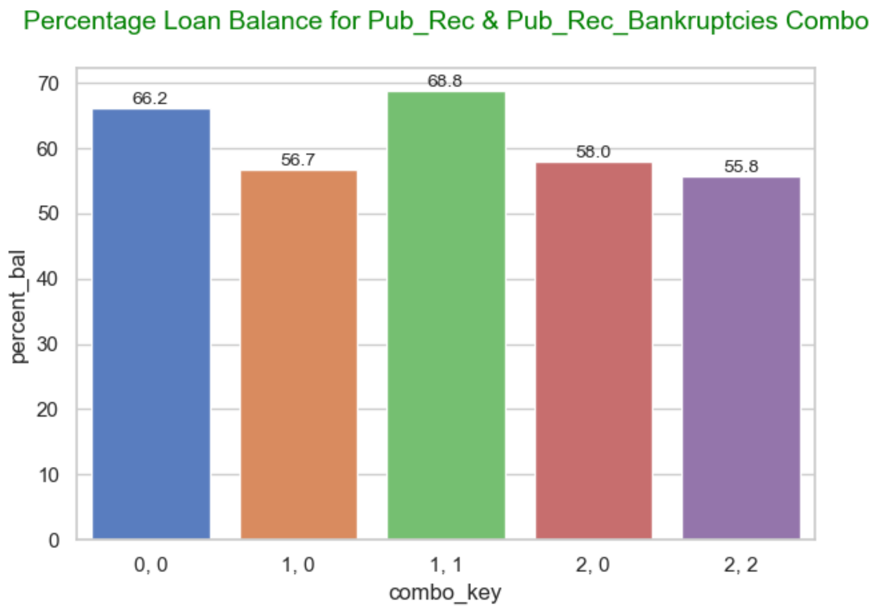

Loan Term by Loan Status

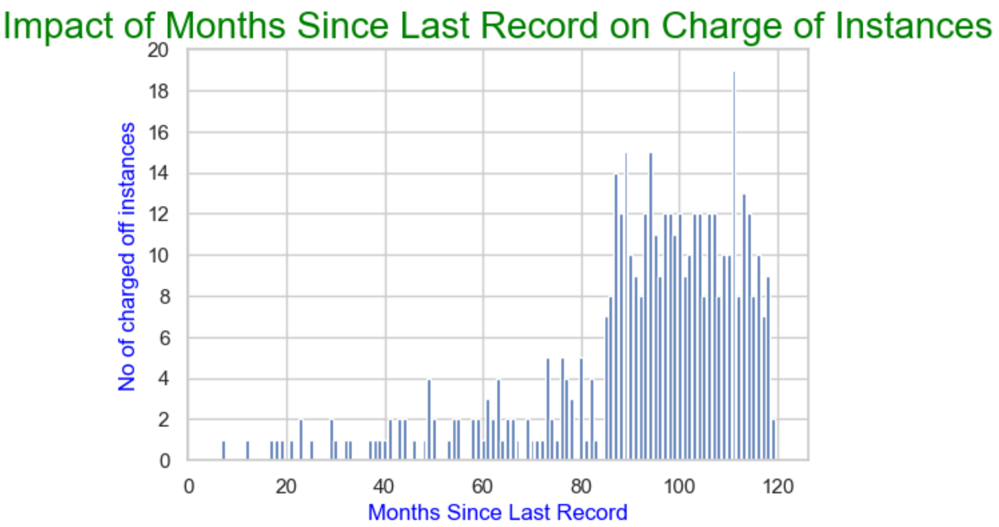

Loan Status Percentages by Delinquancy in Past 2 Yrs

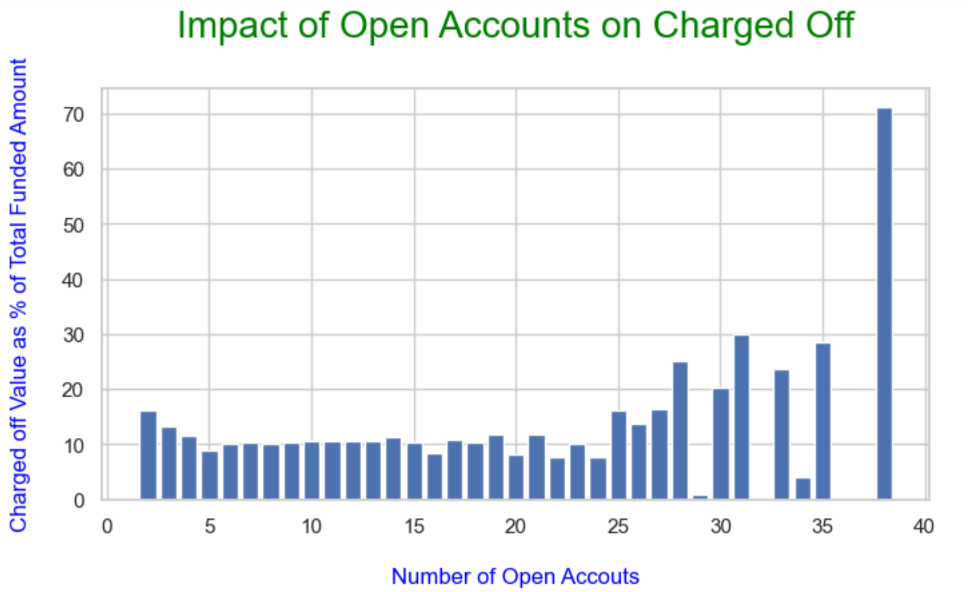

Revolving Balance by Loan Status

- **Loan term, revolving balance** and **delinquency** in last 2 years doesn’t have much impact on Charged-off percentage.

Continuous Variables







- **The impact of pub_rec_bankruptcies has been studied vs. pub_rec and basis the graph one can see that there is no major impact due to bankruptcies.**
- **Hence we will stick with pub_rec in terms of further deep diving.**
- **The open accounts have also been studied in detail and when the open accounts become very high (>25), we see that for those cases the risk goes up**
- **Impact of months since last record is working out opposite to what one might expect. The instances of default are higher in those cases where months is higher. So, we will ignore this variable for further analysis.**

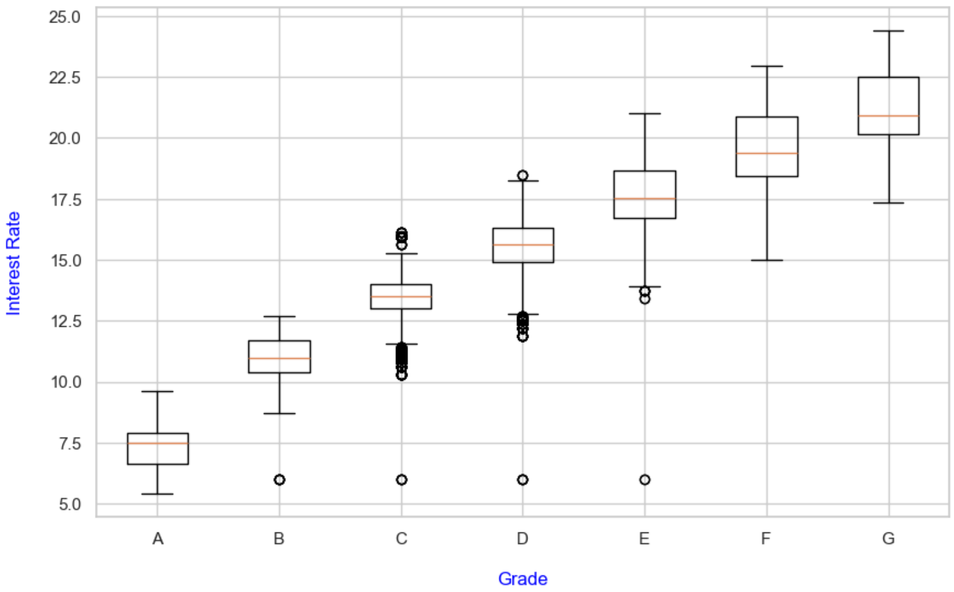
Takeaways

Basis the analysis done in the previous sections, we will focus on the following variables:

- int_rate
- grade
- purpose
- open_acc
- pub_rec
- revol_util

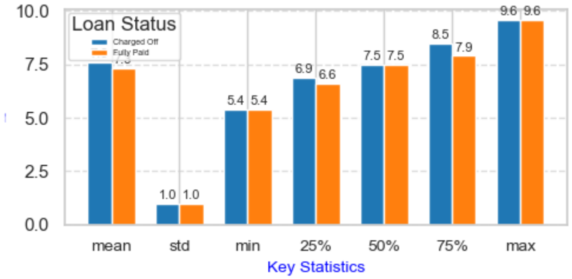
Interest Rate Vs. Grade

Interest Rate Vs. Grade Box Plot

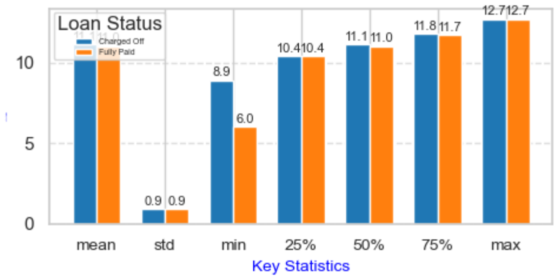


We can see from the above chart that the interest rates charged by the company for lower grade loans are higher, which is a very good practice.

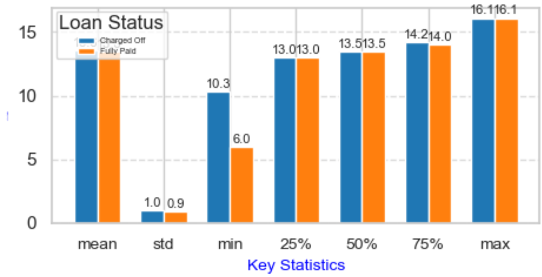
Comparison of Metrics Between Charged Off and Fully Paid for Grade: A



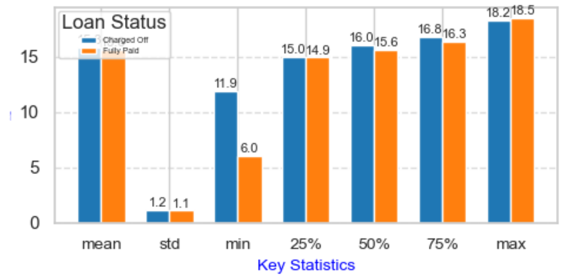
Comparison of Metrics Between Charged Off and Fully Paid for Grade: B



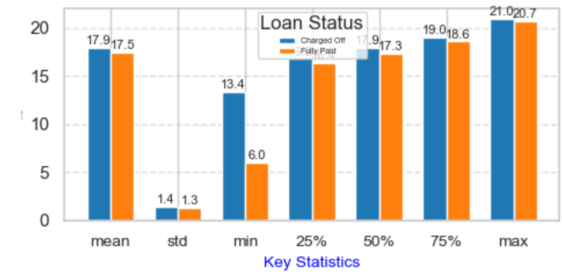
Comparison of Metrics Between Charged Off and Fully Paid for Grade: C



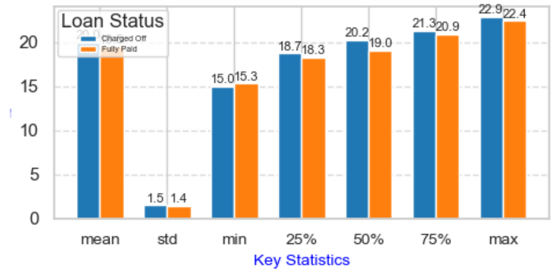
Comparison of Metrics Between Charged Off and Fully Paid for Grade: D



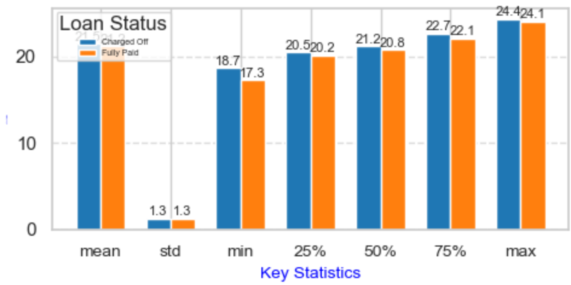
Comparison of Metrics Between Charged Off and Fully Paid for Grade: E



Comparison of Metrics Between Charged Off and Fully Paid for Grade: F

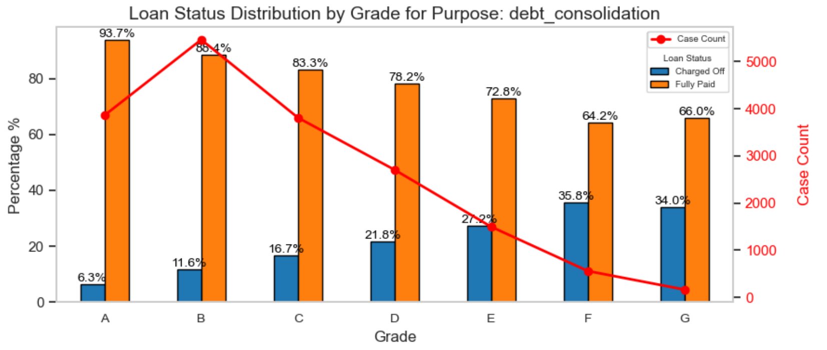


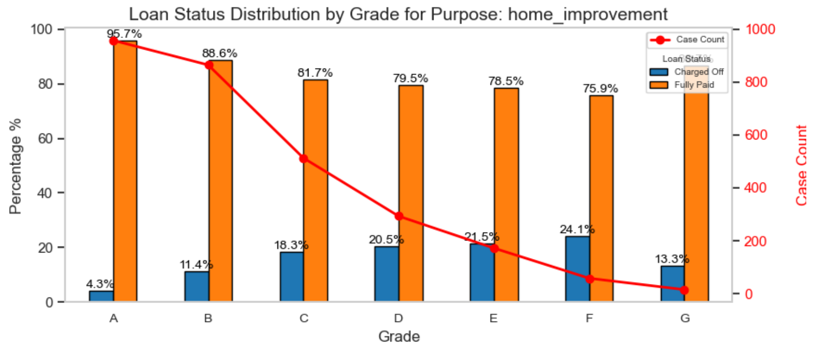
Comparison of Metrics Between Charged Off and Fully Paid for Grade: G

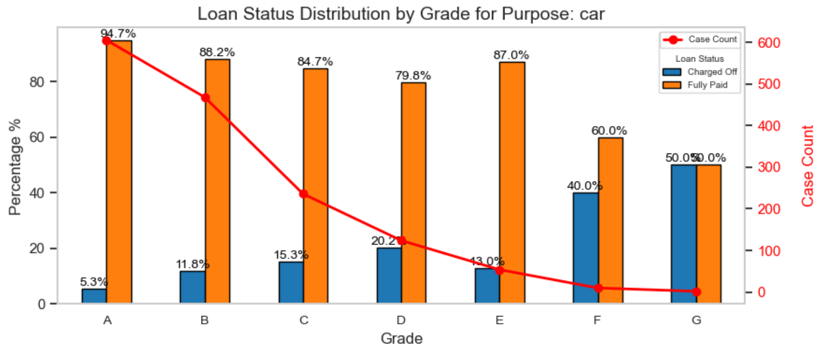


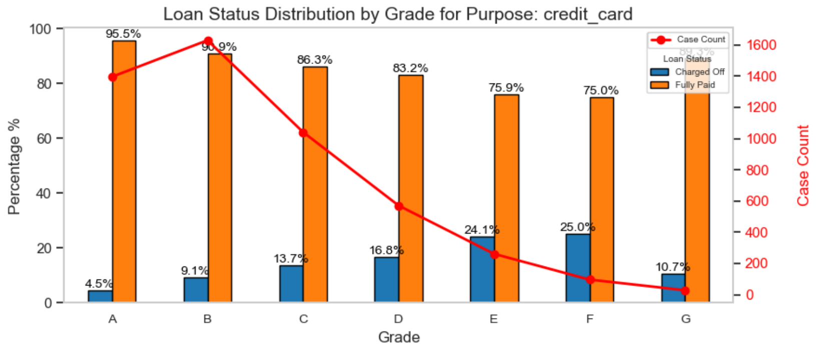
We can see that the interest rates for charged off loans and fully paid loans within each grade does not vary significantly. Hence no impact due to this variable.

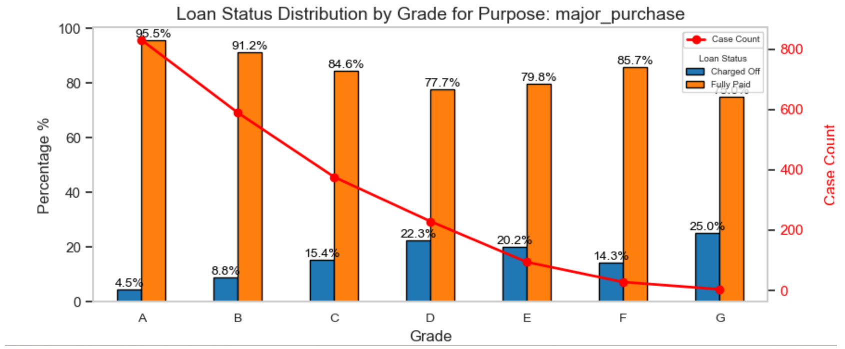
Purpose Vs. Grade

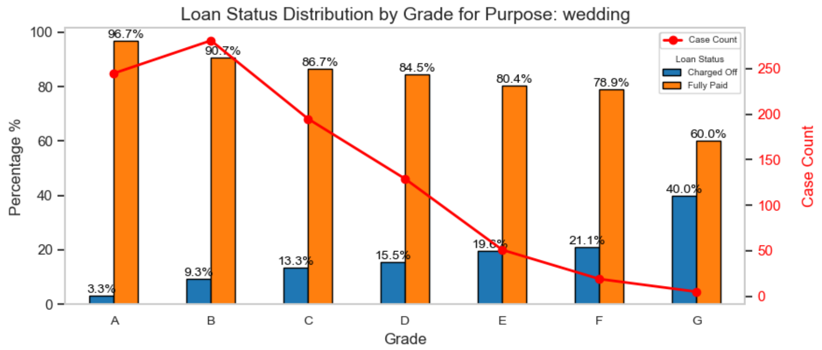

Loan Status Distribution by Grade for Purpose: debt_consolidation

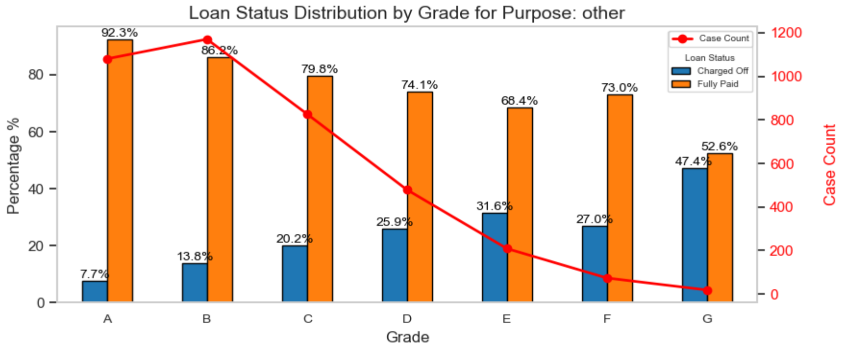

Loan Status Distribution by Grade for Purpose: home_improvement

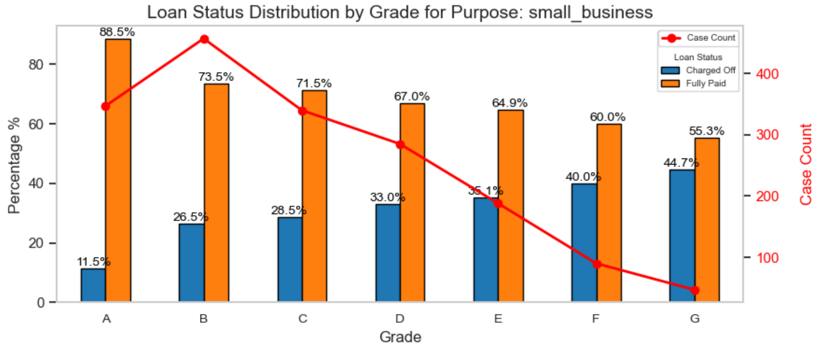

Loan Status Distribution by Grade for Purpose: car

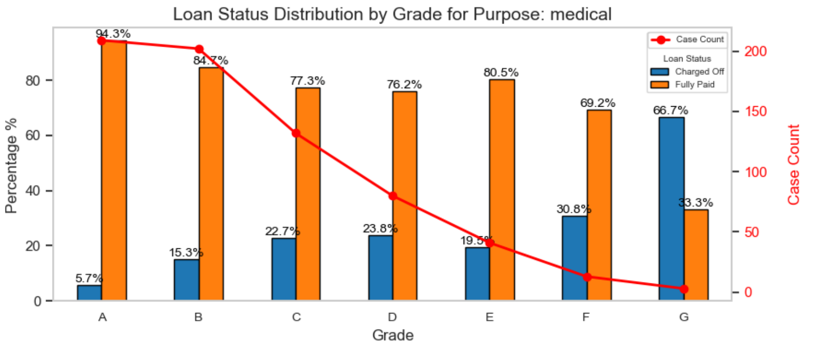

Loan Status Distribution by Grade for Purpose: credit_card


Loan Status Distribution by Grade for Purpose: major_purchase

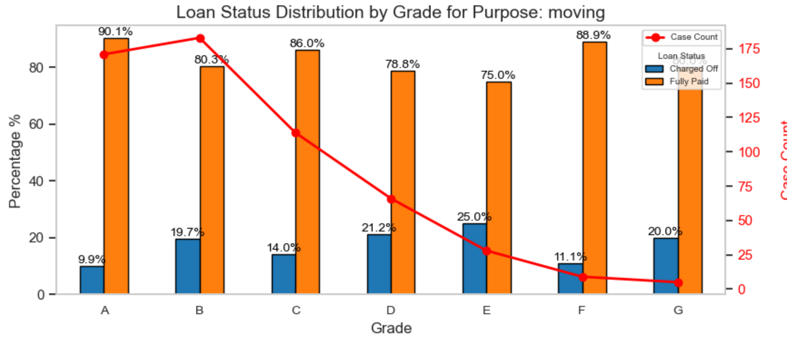

Loan Status Distribution by Grade for Purpose: wedding

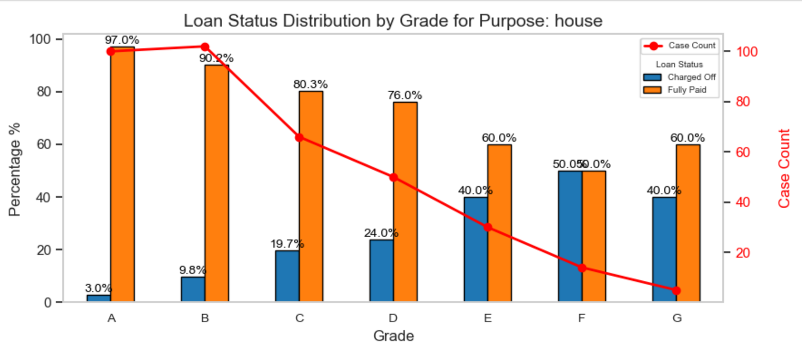

Loan Status Distribution by Grade for Purpose: other

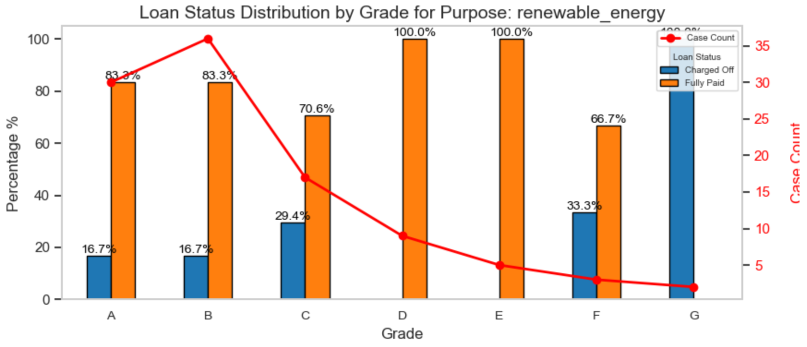

Loan Status Distribution by Grade for Purpose: small_business

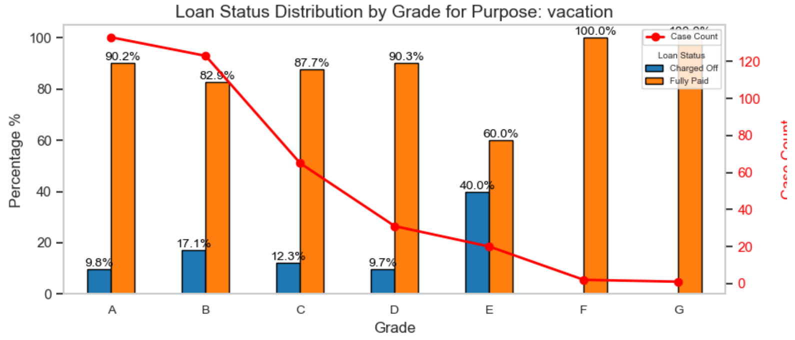

Loan Status Distribution by Grade for Purpose: medical

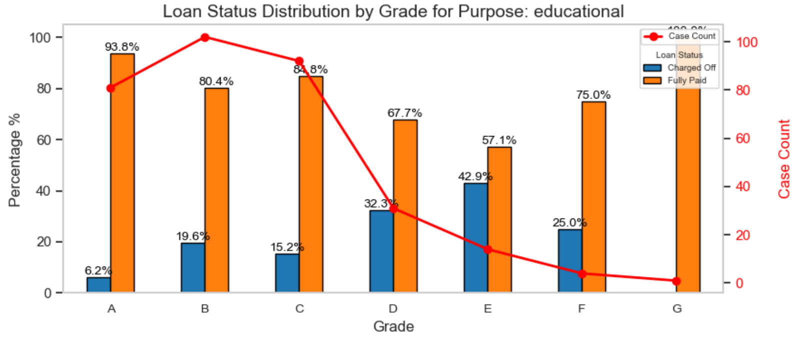
Purpose Vs. Grade


Loan Status Distribution by Grade for Purpose: moving


Loan Status Distribution by Grade for Purpose: house


Loan Status Distribution by Grade for Purpose: renewable_energy

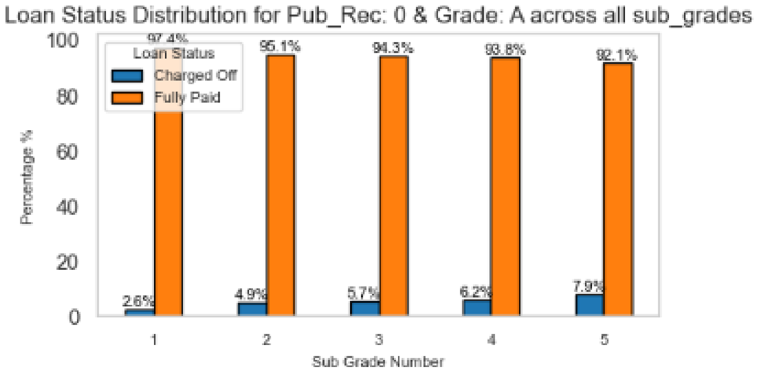

Loan Status Distribution by Grade for Purpose: vacation

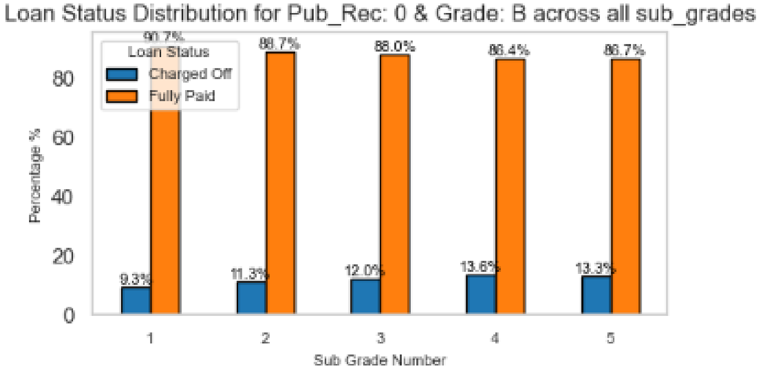

Loan Status Distribution by Grade for Purpose: educational

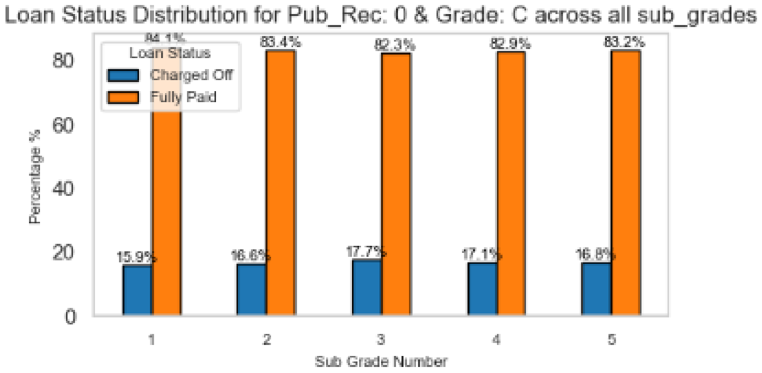
In case of debt_consolidation, small_business, medical, other, wedding, car, house and renewable_energy we find that the charged off % dramatically increases as the loan grade keeps going down.

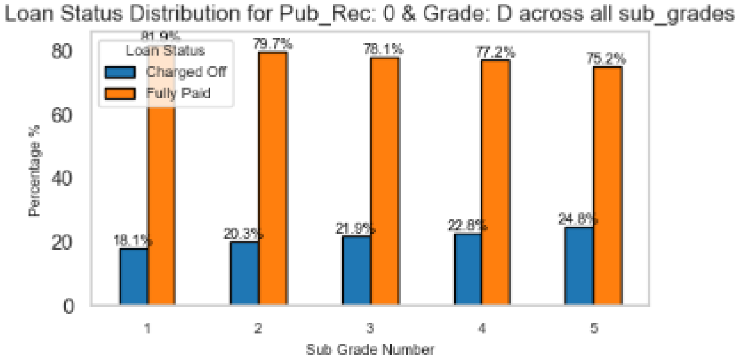
In other cases we are not able to find such a correlation.

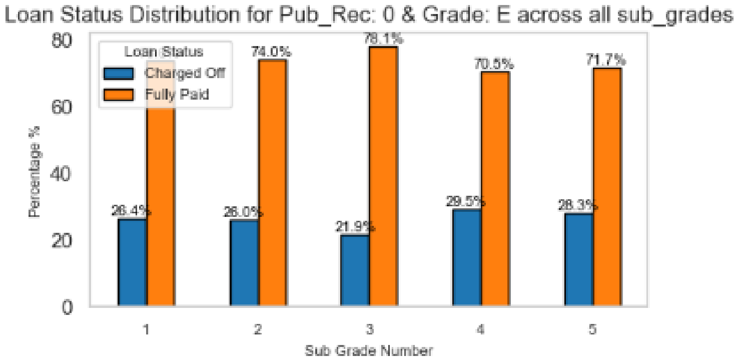
Sub Grade, Grade, Public Record

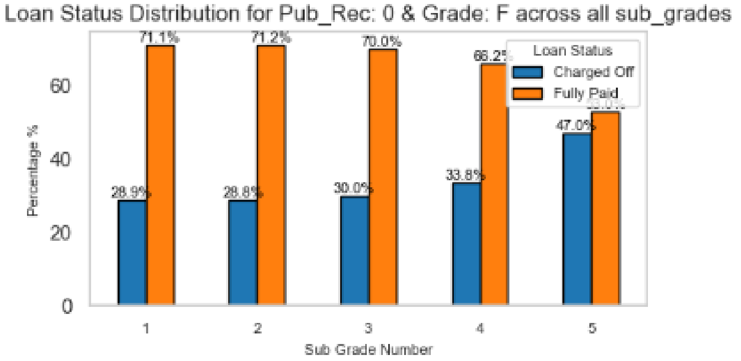

Loan Status Distribution for Pub_Rec: 0 & Grade: A across all sub_grades

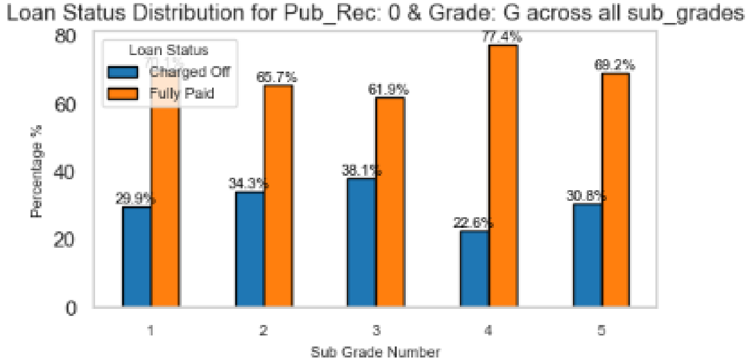

Loan Status Distribution for Pub_Rec: 0 & Grade: B across all sub_grades


Loan Status Distribution for Pub_Rec: 0 & Grade: C across all sub_grades


Loan Status Distribution for Pub_Rec: 0 & Grade: D across all sub_grades


Loan Status Distribution for Pub_Rec: 0 & Grade: E across all sub_grades

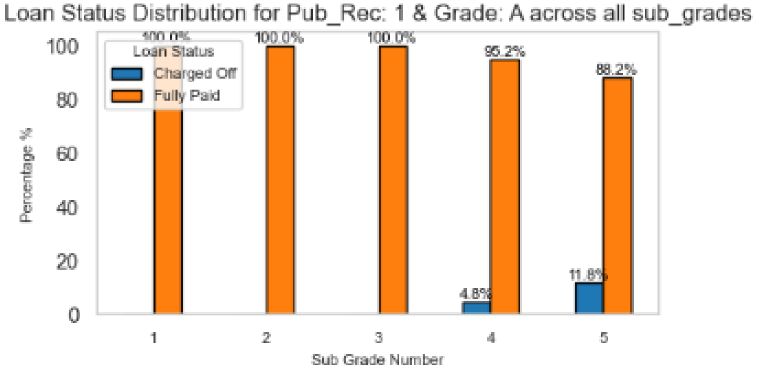

Loan Status Distribution for Pub_Rec: 0 & Grade: F across all sub_grades

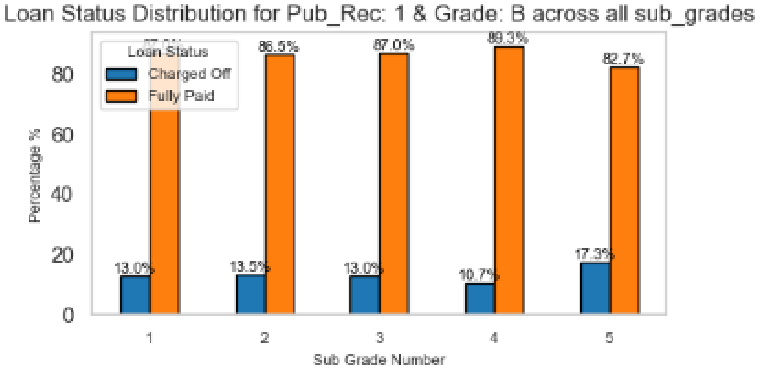

Loan Status Distribution for Pub_Rec: 0 & Grade: G across all sub_grades

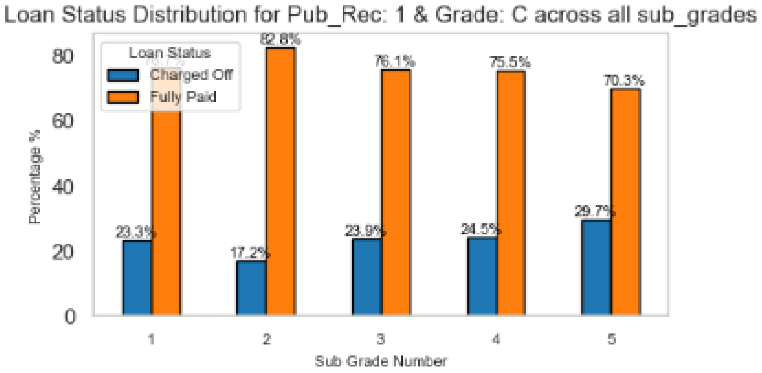
For a given pub_rec 0 and a given grade, the charged off % is not significantly varying with lower sub grades.

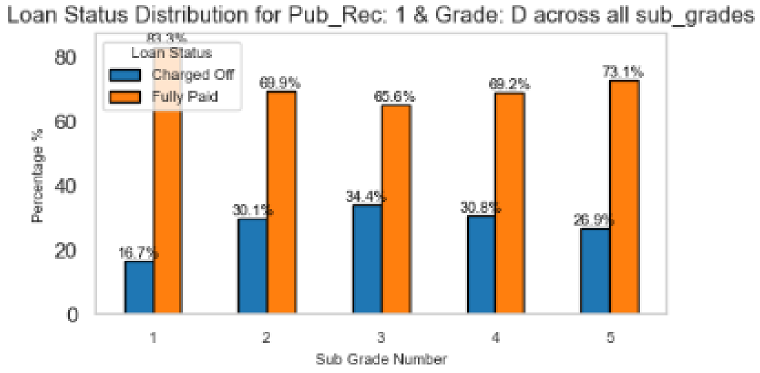
We are noticing one anamoly though in F grade, where the 5th sub grade is having very high charged off %.

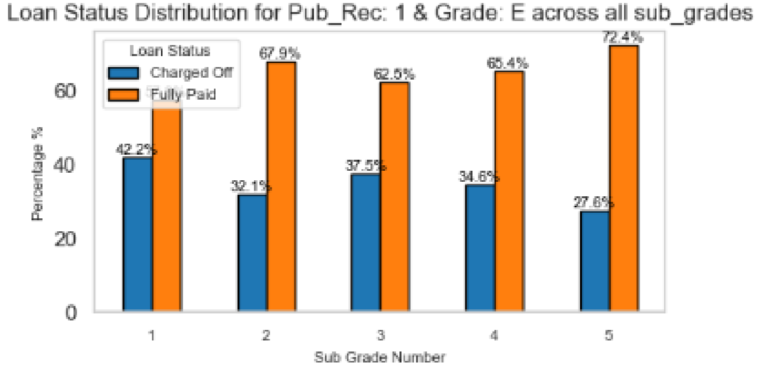
Sub Grade, Grade, Public Record

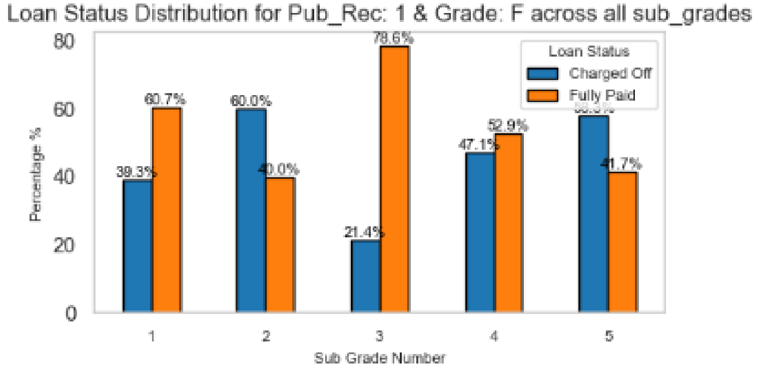

Loan Status Distribution for Pub_Rec: 1 & Grade: A across all sub_grades

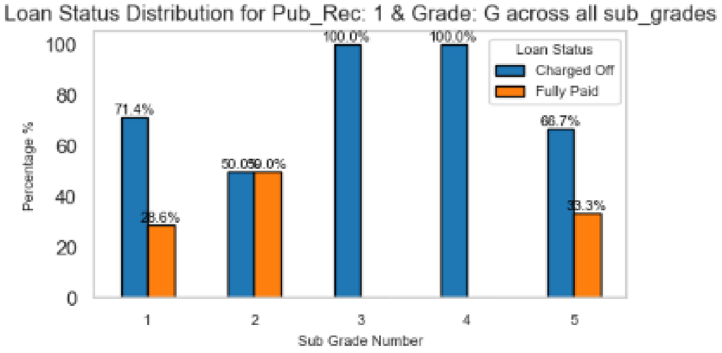

Loan Status Distribution for Pub_Rec: 1 & Grade: B across all sub_grades


Loan Status Distribution for Pub_Rec: 1 & Grade: C across all sub_grades


Loan Status Distribution for Pub_Rec: 1 & Grade: D across all sub_grades


Loan Status Distribution for Pub_Rec: 1 & Grade: E across all sub_grades

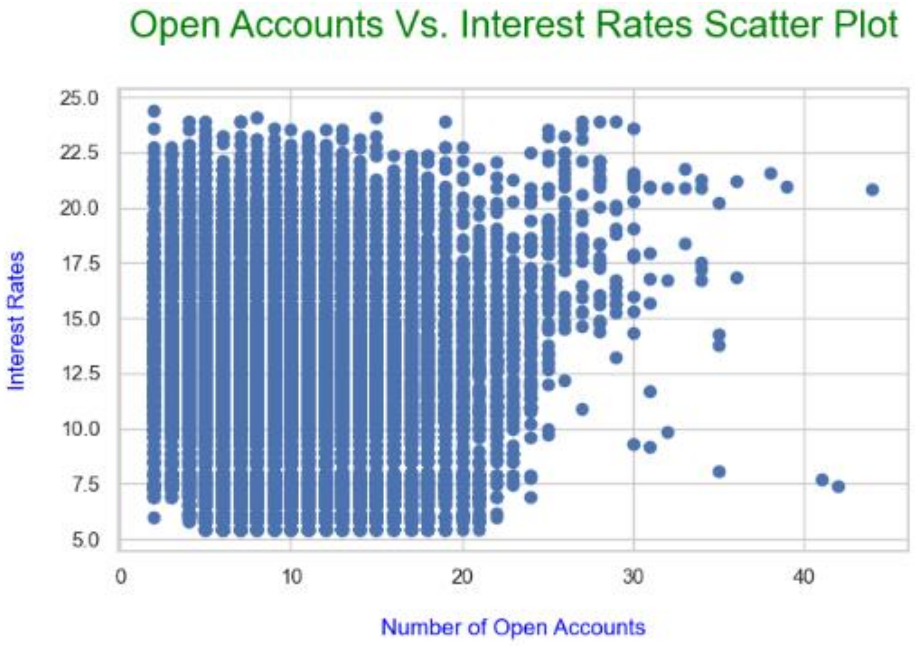

Loan Status Distribution for Pub_Rec: 1 & Grade: F across all sub_grades

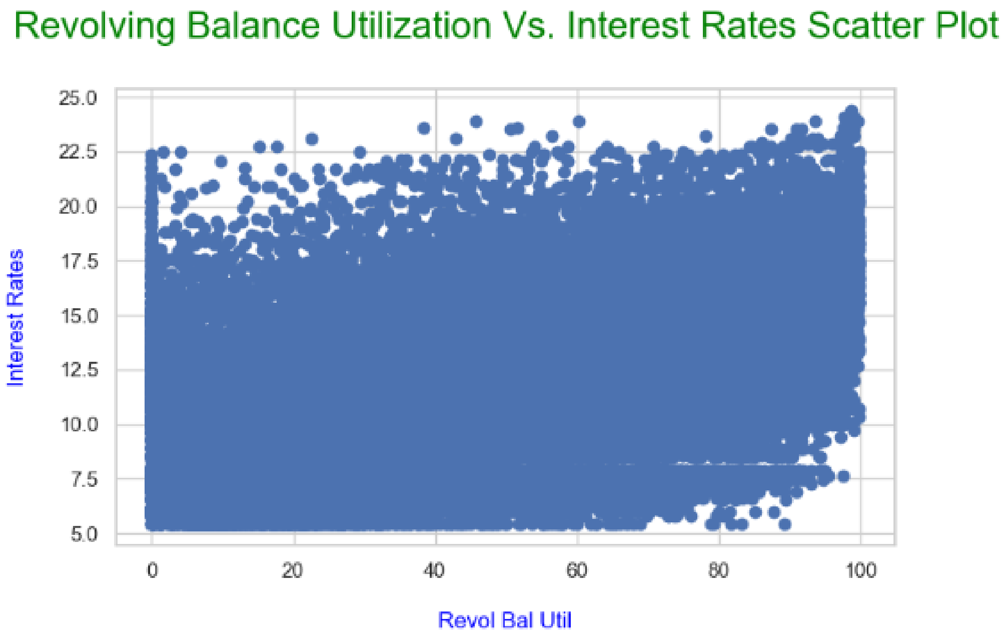

Loan Status Distribution for Pub_Rec: 1 & Grade: G across all sub_grades

For a given pub_rec 1 and a given grade, the charged off % is not significantly varying with lower sub grades.

The movement is quite erratic and we wont be able to associate any correlation with sub grade on loan charged %.

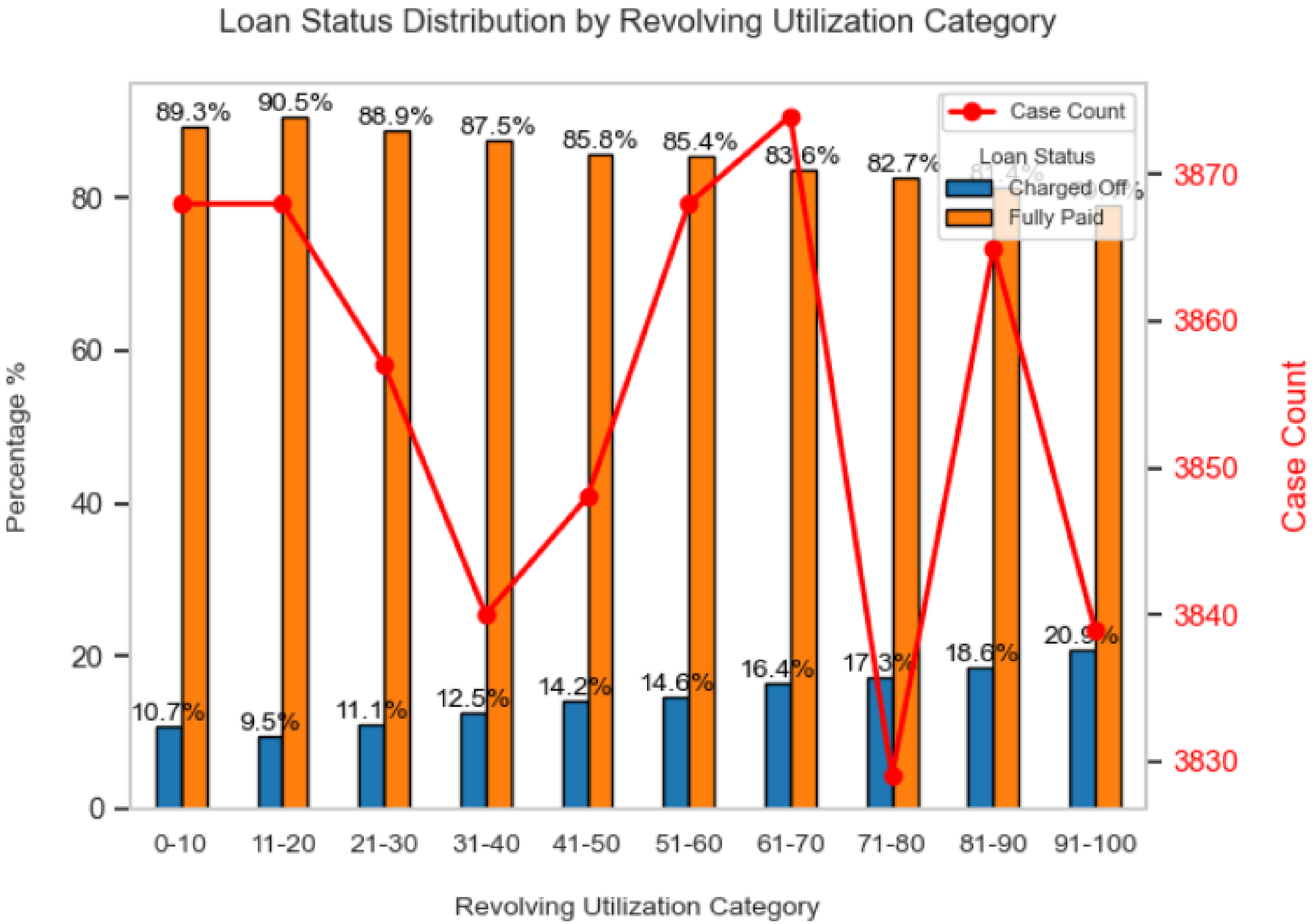
Open Accounts Vs. Interest Rate & Revolving Account Balance Vs. Interest Rate





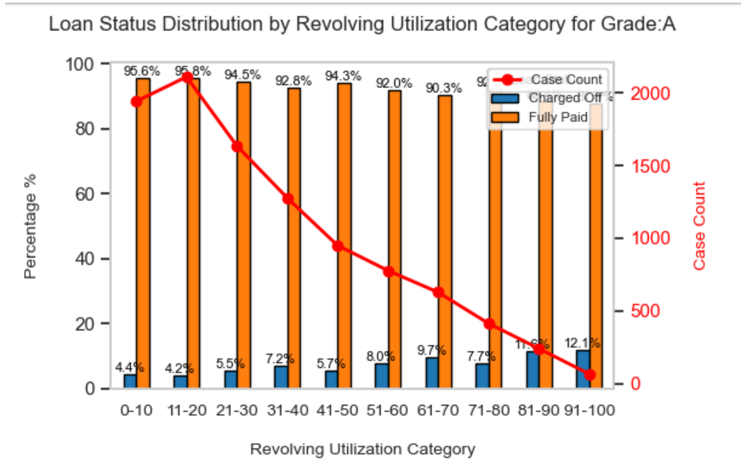
From the above two scatter plots we can see that there is no correlation between number of open accounts and interest rate as well as revol_util_bal and interest rates.

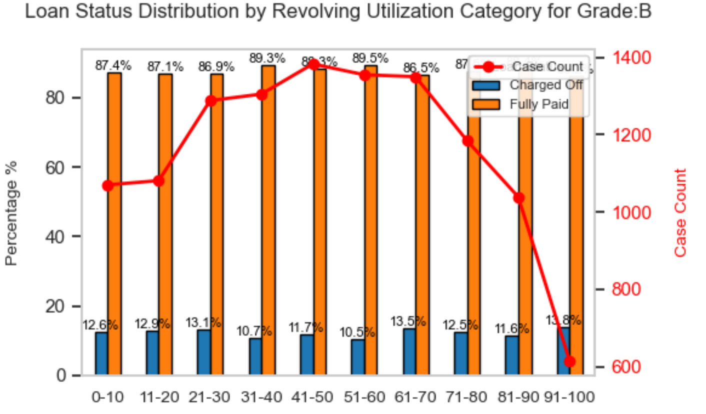
Revolving Balance Utilization Category

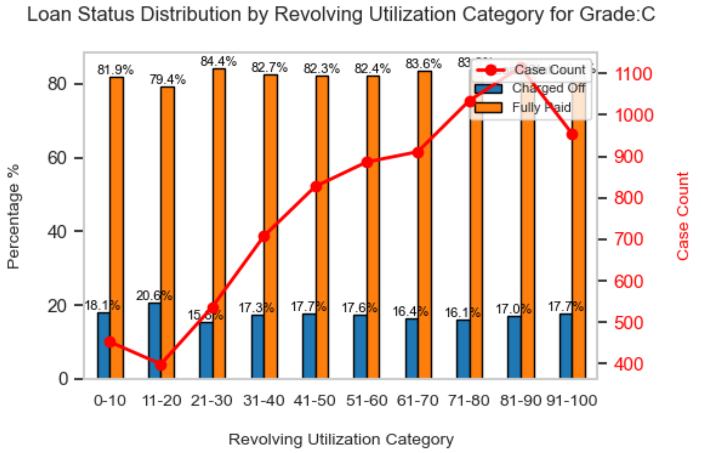


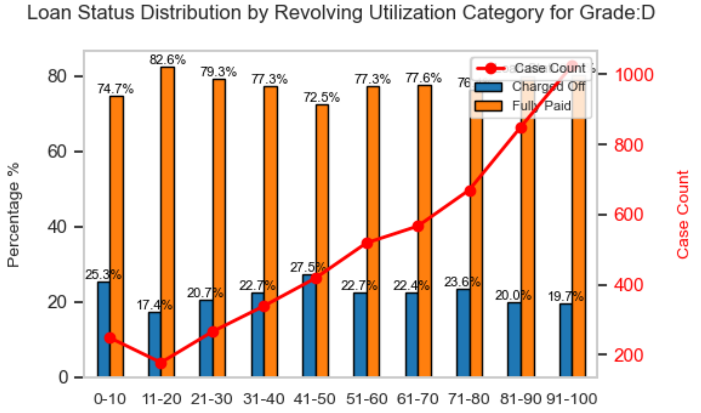
As the revolving utlization % goes up we can see that the charged off% also goes up.

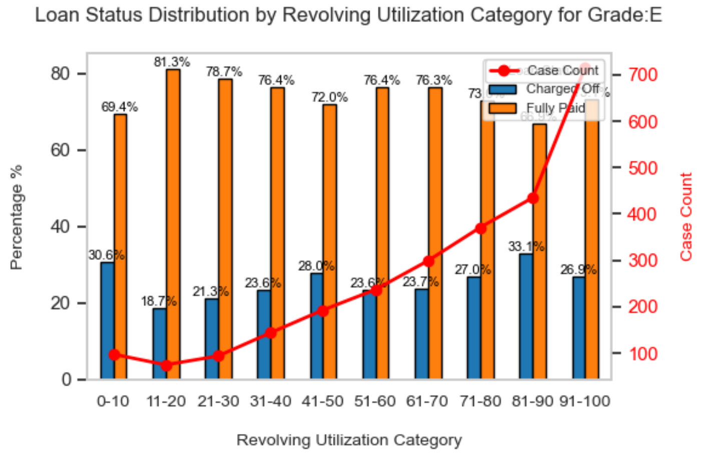
Revolving Balance Utilization Category Vs. Grade

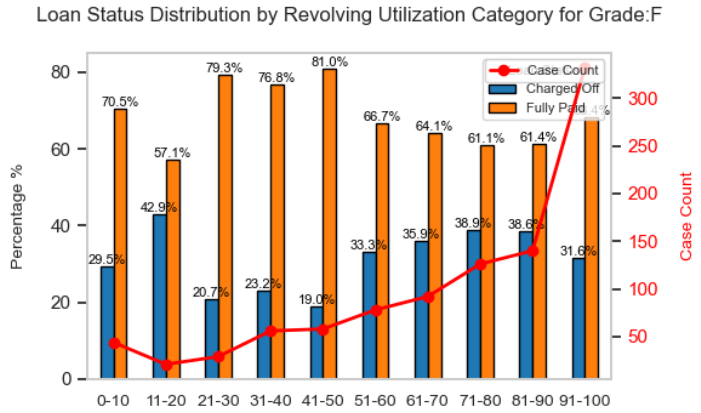

Loan Status Distribution by Revolving Utilization Category for Grade:A

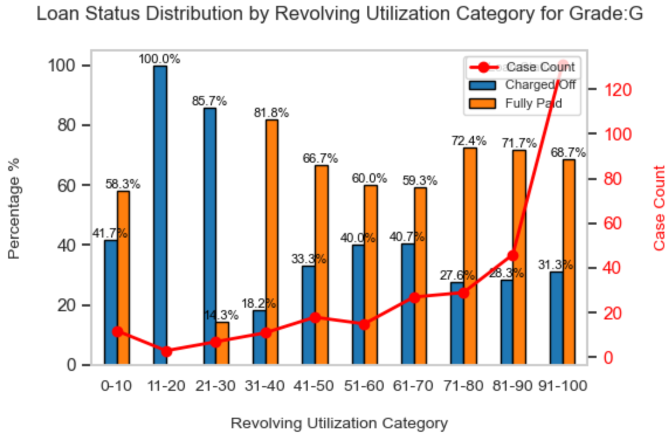

Loan Status Distribution by Revolving Utilization Category for Grade:B


Loan Status Distribution by Revolving Utilization Category for Grade:C


Loan Status Distribution by Revolving Utilization Category for Grade:D


Loan Status Distribution by Revolving Utilization Category for Grade:E


Loan Status Distribution by Revolving Utilization Category for Grade:F


Loan Status Distribution by Revolving Utilization Category for Grade:G

Actionable Insights

Key Takeaways

- Grade is the most important predictor of the loan charged off%.
- Higher number public records is also a very good predictor of loan charged off %.
- Purpose of the loan can also be a good predictor of loan charged off%.
- Purpose along with Grade bi variate analysis can help in further sharpening the insights.
- Very high open accounts can lead to high loan charged off%.
- High enquiries can also be an indicator of high loan charged off%.

Thank You!